SPATIAL MODELING OF ALL-CAUSE MORTALITY

by

Stefan K. Lhachimi

Master Thesis
in partial fulfilment of the requirements for the degree
Master of Science in Statistics (M.Sc.)
March 10th, 2008

submitted to

Prof. Dr. Wolfgang Härdle
Priv.-Doz. Dr. Marlene Müller

Chair of Statistics
Institute for Statistics and Econometrics
School of Business and Economics
Humboldt-University Berlin

Acknowledgements

For completion of this thesis I am heavily indebted to several persons. First and foremost Prof. Dr. Francesco Lagona (Rome) and Prof. Dr. Wolfgang Härdle (Berlin). The former for introducing me to the exciting, and for me novel, field of spatial statistics during my stay at the Max-Planck-Institute for Demography (Rostock). The latter not only for encouraging me to choose this topic but also for being my teacher while studying statistics under his supervision at Humboldt University. Furthermore, Dr. Sigbert Klinke (Berlin) who gave me invaluable advice when I introduced my research question at the statistics colloquium. Finally, I have to thank the R-Sig-Geo email list and here in particular Prof. Dr. Roger Bivand (Bergen) for many quick hints when my S4 class object did not do what I wanted it to do.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Research Question

An event unifying mankind and affecting every single member of it, is the experience of death. However, the exact point of time for an individual cannot be determined in advance and varies strongly in time and place. One branch of demography tries learn more about the determinants of death at the population level. In this, demography always relied heavily on empirical observations through statistical modeling. Recently, the spatial dimension of the collected data has been getting more attention. This is due to pragmatic considerations and theoretical insights.

Pragmatically speaking, a lot of data is only available at some aggregate level. This is often due to the large cost, that collecting data on the individual level implies. But also other reasons play a role, mostly the need or the wish

to preserve a certain level of anonymity of the affected individuals. In some countries – Sweden, for example – mortality data is available giving the apartment number of the individual. In Germany, however, only data disaggregated up to the municipality level (*Kreisebene*) is publicly available. More detailed data, say at the community level, is – when archived at all – only available through special authorization and with huge costs.

Theoretically speaking, it has been increasingly found, that the spatial location of an event can carry a lot of information about the event in question. And that the spatial variation does depend on the distance to related events. This insight might be old for scholars of geography, but the explicit modeling of that phenomenon is a rather new affair for demography. This is – at least partly – due to developments in the field of statistics, that devised methods to deal with spatially occurring data, as assumptions of standard approaches often do not hold. An intuitive explanation for the need for spatial methods lies in the high degree of spatial correlation between region i and its neighbors. The neighbors can be an almost a perfect predictor for region i. Hence, the neighborhood structure has to be modeled explicitly. There is often – but not always – a close the similarity between spatial data and time series data. In time series data the previous observation(s) can be used to predict the present observations, requiring to account for this kind of serial correlation.

The statistical modeling of spatial mortality data aims to test and quantify causal relationships and identify regions with unusual high or low mortality. The former is done to increase the scientific insight into mortality. The latter

can be considered as a form of public health surveillance, assessing whether certain regions have unexpectedly high or low death counts that warrant a closer investigation. Hence, in this thesis we have a research question in two parts. First, we test whether certain independent variables have a significant and sizeable effect on all-cause mortality in Germany, restricting ourselves to the age groups 20 to 49. Second, are there regions with unexpectedly high or low mortality rates (*hot spots* of mortality). A prominent place in our considerations has the idea of an overall spatial trend, e.g. whether the mortality pattern in German exhibits rather an East-to-West trend (ceteris paribus, people in East-Germany are more likely to die) or another spatial trend. But we also caution that spatial modeling can only yield partial insights due to the limitations of the data and techniques (e.g. ecological fallacy).

1.2 Outline

The thesis is divided into three main chapters: *Mortality*, *Spatial Statistics*, and *Analysis*. The rationale for this somewhat static division is to enable the reader to skip parts that are of marginal interest to him or her and concentrate on the topics of his or her choice.

In section 2.1 we elaborate on the main goals of spatial demography, mainly the modeling and testing of – often easier – available data and the identification hot spots. We then proceed giving a brief account of some of the main determinants of mortality. These are mainly the socio-economic well being,

the endowment with public (health) infrastructure and the effect of migration on the population composition. We pay special attention to the role that a spatial trend plays as a measurement of absolute regional difference. In section 2.3 we discuss several approaches of measuring mortality, by no means a trivial task. The measure has to be standardized to account for several factors – in particular age composition – to be truly comparable between different populations or regions, respectively. We give reasons why the standard mortality ratio (SMR) is to be preferred for the task ahead and why we do not employ life expectancy.

In chapter 3 we give an overview of the methods used for our analysis. We start out with placing our research question within the general model of spatial statistics showing that for our kind of data the lattice data approach should be chosen. In section 3.2 we discuss some of the different existing specifications of the neighborhood structure that – in the case of lattice data – carry the information about the spatial structure. In section 3.3 we briefly explain the basic mapping of statistical variables and the choices one faces. In section 3.4 we introduce the standard statistic – Moran's I – to quantify and test for spatial autocorrelation. This measure shows the existence and the strength of the spatial association within lattice data. In the last section of this chapter (section 3.5) we introduce the spatial error model as one modification of of the standard regression approach. We show that need for explicit accounting for the spatial nature of the data and how the spatial error model is estimated.

In chapter 4 we finally conduct the empirical analysis. We map and describe

the dependent variable – log of SMR – and establish its suitability for the used regression specification. We do so – inter alia – by showing the high degree of spatial association it exhibits. After describing the choice of spatial weights and the independent variables, the regression results are presented. In the last part we map the identified hot spots. First we just use the 'plain' dependent variable, the we use the residuals of the regression model that have the advantage to be corrected for the spatial structure.

In the appendices we give further information about the analysis. In section A we give choropleth maps of the dependent variables for further study by the reader. In section B a table with all the regions used in the analysis is shown. It contains also the number of observed deaths and the calculated SMR. In section C the main R-libraries that have been used for the analysis are acknowledged.

Chapter 2

Mortality

The theory of mortality tries to identify determinants of change in life span and future developments. At this point in time there does not exist a *single* theory of mortality, rather a set of observations and explanations which derive mostly from two sources (compare for the following Preston et al. (2005); Cutler et al. (2006)): empirical findings and mathematical modeling. Before the 18th century average life expectancy did not increase much. However, since then mortality increases slowly but steadily and has now reached an all time high of 85 years for Japanese women. Several factors have been suggested explaining the mortality decline; these are improved nutrition, better public health, better medical knowledge, and behavioral change. The increase in food production does not only reduce the occurrence of famine, but undoubtedly makes people less likely to die from infectious diseases. Better public health mostly translates into better sanitation in the urbanized areas in the 19th century but also into

the drainage of swamps, pasteurizing of food, and better housing. The increase in medical knowledge – treatment of disease and better prevention by vaccines – are often the reason that comes to mind first. However, the mortality decline started already before the vast innovations in medical science occurred. Lastly, the changed behavior of individuals has to be mentioned. Not only do people have to adhere to a healthier lifestyle – such as less alcohol, no smoking, balanced diet, personal hygiene – they also have to cooperate with medical treatment facilities. There is evidence, for example, that people that have more trust in medical insight are more likely to utilize medical facilities to their advantage. The actual ranking of these factors, however, is still under discussion.

2.1 Aims of spatial mortality analysis

Two main goals exist in the analysis of the spatial dimension of (all-cause) mortality (Lagona and Barbi 2006): the modeling of external *determinants* of mortality and the identification of *hot spots* of mortality. The spatial modeling of external determinants of mortality is often necessary due to the fact that some determinants of mortality are only measured or reported at the aggregate level – such as occurrence of medical facilities. Furthermore, the spatial variation within a country might indicate different health-relevant behavior of the studied population. For example, diet and alcohol consumption habits differ often regionally. Moreover, the regional variation in external factors – most

notably climate, but also income or education disparities – may have an effect on the risk to die.

Within a country clusters of unusual high or low mortality can be observed. Initially, it has to be determined if these cluster are – statistically – significant or just occurred randomly. A hot spot (defined as an area with unusual small or large mortality) often, but not necessarily, contains valuable information for the future study of determinants of mortality. For example, recent research findings showed unusual high levels of longevity on some parts of the island of Sardinia which are now being studied closer by analyzing individual data (Caselli et al. 2002). On the other hand, persistent clusters of high mortality indicate the existence of environmental or behavioral factors that warrant the need for public health interventions.

Aggregate (mortality) analysis, however, is cursed with the problem of *ecological fallacy* – meaning that relationships analyzed at the aggregate level may lead to different conclusions when compared to the analysis of individual data (Greenland 1992). A famous – albeit gross – example is the correlation between illiteracy rate and foreign born population in the United States. At the state level (share of illiterate and share of foreign born), the correlation is .11; whereas at the individual level (being illiterate and being foreign born) the effect changes sign and has a correlation of -.53. Certainly, using individual-level data is the gold standard in every statistical analysis of human behavior. However, such data is often not available and one of the main goals of statistical modeling is to use the best method available to extract the most information

from the given data. Furthermore, the problem is mitigated by the use of models where the causal path is well established at the individual level.

2.2 Spatial Differentials in Mortality

At the population level several factors are suggested to explain the different mortality levels. In our exposition we restrain ourselves to four of the most accepted factors (Preston et al. 2005; Cutler et al. 2006). Those are the *socio-economic gradient*, the *medical infrastructure*, the effect of *migration*, and absolute *spatial differences*.

Socio-Econmic Gradient

The socioeconomic gradient[1] points to the observation that a higher socioeconomic status leads a lower mortality. A problem in this field is to identify the causality. It could be that higher socioeconomic status leads to better health and lower mortality. Or is it that people, who already have a better health - or a better genetic makeup – are able to obtain a higher socioeconomic status. A further problem with this line of research is that socio-economic status remains a somewhat fuzzy concept. Everybody has an intuitive understanding of what is conveyed by it, however, it is difficult to find a reliable and universally accepted measure. But differences in socio-economic status remain whether income, education, or occupation is chosen as an indicator. It is not

[1]The term is used as there exist *graded differences* in health across ranked groups

an absolute income level which is important, but the degree of inequality in a given group. It is notable, that male mortality shows an higher correlation to socioeconomic status than female mortality.

At the country level, the mortality differential between rich and poor countries is well established. In rich countries less than 1% of deaths are among children whereas in poor countries it is up to 30%. In rich countries most people die from cancers and cardiovascular diseases; in poor countries deaths are still caused by infectious diseases. Although an increase in life expectancy in poor countries has been observed since the Second World War the AIDS/HIV epidemic has offset these improvements since the middle of the 1980ies.

But also within a country regional disparities of life expectancy can be witnessed. A well studied example for this is the United States where. The average male resident of Baden-Wuerttemberg lives on average 1.3 years longer than the average resident of Germany (in 2000). This effect is slightly weaker for females, but still exists. The overall life expectancy in east Germany, despite closing the gap to the west since the reunification, is still 1.6 years lower (in 2000) for males than in the national average (Cromm and Scholz 2002).

Public Infrastructure

Public infrastructure, in this respect, consists mostly of public health measures such as sewage systems and public health care. In particular, the almost universal access to health care is – together with better nutrition and increased

hygiene – one of the main reasons for the observed mortality decline in the developed world. The disparities in spending and maintaining for such infrastructure between regions is often able to explain (at least) partly the differences in mortality. This is sometimes dubbed as the *medical underspending* hypothesis.

In Germany, regional variations in the endowment with such infrastructure exists to a certain degree. The municipalities – in the context of local self-government – are responsible for establishing and running of hospitals. Certainly, the state level (*Länder* exercise an influence by supporting the communities). But in the very end the final decision lies in the responsibilities and capabilities of the respective municipality.

Migration

Another important effect on spatial mortality differential is certainly the so called *healthy migrant* effect. This term describes the phenomenon, that for international migration a selectivity process conditional on health can be observed. This is certainly true for organized international immigration, where the receiving countries did and do extensive health checks before admitting immigrants, sending back individuals deemed too frail. The medical examination on Ellis Island in the American case or the examinations in Turkey for the guest worker program in the 1960s by German doctors are vivid examples of this.

For the voluntarily migration within a country the effect is less clear cut. First, the movements are not legally restricted based on the health status of an individual. Second, a healthy person might be more able to find work in his own locality therefore, ceteris paribus, reducing his or her propensity to migrate. On the other hand, a more healthy migrant is more likely to find work in prosperous region than an ill or handicapped person. Hence, for internal migration the effect could be both ways, high in-migration levels could increase or decrease the observed mortality differential.

Absolute Regional Differences

Finally, the last category of explaining factors within the field of spatial mortality analysis are *absolute regional differences*. Many (relevant) variables vary systematical through space. A good example of this is the the effect of climate and accordingly temperature on mortality (Rau 2007). Furthermore, some individual behavior is culturally influenced and hence factors such as smoking-behavior or alcohol consumption vary by region. In general, the differences of (regionally and culturally affected) diet might greatly contribute to the explanation of regional variability on mortality. On a larger scale, the genetic make up of a population has an effect on mortality as well. For example, in the United States some differences in population health outcomes can be explained with the difference in the ethnic population composition.[2]

[2]The higher incidence of stomach cancer in the Midwestern United States, for example, is due to the preference for smoked fish by the descendants of Swedish immigrants who mainly settled in this region in the 19th century.

The problem of such factors that vary systematically through space is twofold. First, often these variables and their effect are not known or are not fully understood. Second, these variables are often difficult to measure. In particular when conducting analysis at the aggregate level. Hence, these factors are often lumped together in an overall spatial trend. This is to a certain degree a residual category, trying to account for all spatially varying influences. A very good example for that is the North-South trend within Italy. Despite controlling for a series of variables, still a significant disparity exist (Lagona and Barbi 2006).

2.3 Measuring Mortality

Finding a (population-based) measure of mortality is hampered by two demographic properties that have to be taken into account. The *populations size* and the *population composition* (compare here and in the following Chiang (1984)). The most simple and somewhat meaningful measure of mortality is the *crude death rate* (CDR). The CDR is calculated by the number of recorded deaths in a given time period divided by the number of person-years in this time period:

$$\text{CDR} = \frac{P}{D} = \frac{\sum_a D_a}{\sum_a P_a} \tag{2.1}$$

The person-years P are calculated by multiplying the number of persons in the given time period, say one calendar year, by the time they were alive in

this time period. For most people this is then one year. However, if somebody exits or enters the population – say, through death or migration – only the time fraction he was present in the country or region in question is used in the numerator (with other words: the *event* is divided by the *exposure*). But such information is usually not available, only annual changes are reported. The person-years are then approximated by assuming that all deaths occur precisely in the middle of the time interval $[0, T]$. This assumption is for short time periods – one year – and most age groups reasonable and therefore widely used. Hence, if P_a is used, we always mean the person-years based on the mid-year population for the age group a.[3]

Table 2.1: Comparison of CDR for two populations with different age structures: P_a, D_a, and m_a are the population, number of deaths, and the deaths per 1,000, respectivly for age group a (Chiang 1984)

	Region A			Region B		
	P_a	D_a	m_a	P_a	D_a	m_a
Children	10,000	80	8.00	25,000	250	10.00
Adults	15,000	165	11.00	15,000	180	12.00
Seniors	25,000	375	15.00	10,000	160	16.00
Total	50,000	620	12.40	50,000	590	11.80

A problem of the CDR is, that it does not take into account the age structure of the population. As age is the best predictor for the mortality risk of an cohort a measure neglecting age is flawed and only partly useful. A fictional,

[3]In general age groups should not be made to small to avoid larger variability due to the rareness of the event. For example, in in 2006 in the whole country of Sweden no female died in the age of 7. A single occasion, say a car accident of twin-sisters or fire in a elementary school would distort the rates dramatically. The WHO recommendation – depending on the age in question – is to use 5-year age intervals (Shryock and Siegel 1988)

yet instructive example is given in Table 2.1. For every age group in region A, the age-specific death rate m_a is lower than for region B. Nevertheless, the CDR is lower in region A. This is due to the higher share of *Seniors* in region A than region B. This phenomenon can be observed very often in practice, in particular if small regions are analyzed.[4]

One solution is just to report the mortality for short age groups separately. This, however, would offset the goal to find a *single* measure for mortality. Therefore *age adjusted* – or *population age structure adjusted* – mortality measures are needed. In the following we introduce, and discuss in turn, two different standard approaches. These are namely *direct age-standardization* and *indirect age-standardization*. In the last subsection we give an explanation why we are not using the life expectancy as a measure.

Direct Standardization

To compare the morality regime of two or more regions with different age structures an external standard is needed, the so-called *standard population* (denoted by the superscript s). The *direct method death rate* (DMDR) is thus a weighted mean of the age-specific death rates of the region(s) in question (denoted by the superscript u) applied to the standard populations age proportions:

[4]But it can also be witnessed for larger countries, when the age distribution differs dramatically. Compare, for example, the Swedish and Kazahk females in 1992. Although Sweden has the lower death rate in every age group, the overall CDR is higher due to the comparatively older population of Sweden(Preston et al. 2005).

$$DMDR = \sum_a \frac{P_a^s}{P^s} m_a^u = \frac{\sum_a P_a^s m_a^u}{P^s} \tag{2.2}$$

The numerator $\sum_a P_i^s m_i^u$ are the number of deaths that would occur in the standard population if it would be exposed to the mortality regime of a given region. The main aim of the DMDR is to compare two or more regions with each other. It eliminates the difference in age composition of the regions in question, but this comes with a price. The measure is now a function of the age composition of the standard population. Depending on different choices of the standard population this might lead to contradictory results when comparing regions with vary different mortality patterns (see Table 2.2).

Table 2.2: Comparison of DMDR for two populations with different age structure: P^s standard population, m_a deaths per 1,000 , and D_a number of deaths for age group a Chiang (1984)

		Region A		Region B	
Age group	P^s	m_a	Expected D_a	m_a	Expected D_a
Children	35,000	8.00	280	10.00	350
Adults	30,000	11.00	330	12.00	360
Seniors	35,000	15.00	535	16.00	560
Total	100,000		1,145		1,270

Indirect Standardization

One problem for the DMDR is that for small regions or small age groups the age-specific death rate m_a is not stable over time or difficult to obtain. The *indirect method death rate* (IMDR) overcomes this problem by using the death rates of the standard population applied to the population of the region(s) in

question. This figures are more reliable, since they are either based on a larger population or are simply taken from existing standard tables.[5] The IMDR is calculated by multiplying the CDR of the region u in question by the ratio of the CDR of the standard population s if it had its own mortality rates (m_a^s) but a age structure like region u:

$$IMDR = \frac{\frac{D^s}{P^s}}{\frac{\sum_a P_a^u m_a^s}{P^u}} \left(\frac{D^u}{P^u}\right)$$

When the age structure of region u and the standard population s are equal, then the first factor in equation 2.3 becomes unity. In this case the IMDR equals the crude death rate of the community.

The IMDR can be further simplified to lend a very useful measure, the *standard mortality ratio* (SMR). A closer look at the IMDR reveals that the only quantity depending on region u is the ratio:

$$SMR = \frac{D^u}{\sum_a P_a^u m_a^s} = \frac{\text{No. of Observed Deaths}}{\text{No. of Expected Death}} \tag{2.3}$$

The SMR can be interpreted now in a very useful way. The numerator can be seen as the number of *observed deaths* in region u. The denominator is now the number of deaths we *expect* when the mortality regime of the standard population s would be in operation. A SMR larger than 1 means that the mortality in the region in question is higher than in the standard population.

[5] The standard collection of such *model life tables* for different mortality regimes are compiled in Coale et al. (1983).

A SMR smaller than 1 means that the mortality in the region is lower than in the standard population.

This makes this measure now ideally suited for the analysis of regional mortality within a country. The national mortality rates are now taken as the mortality rates of the standard population m_s and applied to the age structure of the respective regions of interest. Now it can be easily classified whether a (sub-)region has a higher or lower mortality than the country as whole, without running into the problem that the age-specific death rates in this region might be unstable due to small population numbers. For the purpose of statistical modeling, often the natural logarithm of the SMR is taken. These is called the *log SMR* or *log relative risk* (Pocock et al. 1981; Lagona and Barbi 2006).

Life Expectancy - useful but difficult

We refrain from using life expectancy as measure. Surely, life expectancy has some nice appealing properties as a measure of mortality, namely it is independent of the age structure of the population and it is measured on a continuous scale allowing the use of a host of well established statistical methods. However, some fail to recognize that is a very sensitive measure: it demands a lot from the existing data as it is calculated using a period life table (Vallin and Caselli 2006).

The following criticism should not be understood as a general rejection of life expectancy as a measure. It is the gold standard in mortality analysis. However, in practical work – like for this thesis – it is often difficult (or

almost impossible[6]) to get the high quality data that is needed. But using sub-standard data would render the measure almost useless as the following discussion will show.

In a cohort life table as many age groups as possible should be included. Unfortunately, for the oldest age group, most statistical offices report only one age group: *75 years and older*. Due to the increase in longevity, however, this is not sufficient and more age groups in higher ages are needed. In particular in the light of recent research that shows that the mortality of the oldest old is actually declining again (Thatcher et al. 1998), rendering the usual approximations at least questionable. This is often called the *table closure* problem.

Another problem is the first year age group. For all groups the assumption is made that the death are uniformly distributed throughout the year. This assumption, however, is not reasonable for the first year age interval (e.g. newly born infants). Most die within a few day or weeks of their birth. But this effect is not stable in cross-country comparison as it depends heavily on pre- and post-natal medical care making a widely accepted approximation very difficult and therefore requiring more accurate mortality data (such as the neonatal death rate and the post-neonatal death rate). This is often dubbed the *infant mortality* problem.

[6]Due to data privacy regulation in Germany it is very difficult to get certain data at the local level at all.

Chapter 3

Spatial Statistics

The origins of spatial statistics are rather old and the first use of spatial data can be traced back to Halley in 1686 who analyzed trade winds and monsoons using a map of land forms. A more explicit form of spatial modeling was undertaken by Student in 1907 who studied the distribution of yeast cells by dividing the surface into 400 squares. He discovered that the cells followed a Poisson distribution. R. A. Fisher, who pioneered statistical analysis by, inter alia, conducting agricultural experiments which obviously have a spatial dimension. He did not explicit model the spatial nature of the experiment, on the contrary he did develop techniques to neutralize this effects; by that showing that he was aware of its existence.

Since the mid-20th century spatial statistics started to developed as a field in its own right. In his comprehensive treatment of spatial statistics, Cressie (Cressie 1993) suggests a general statistical model for the analysis to accommo-

date the different classes of spatial data. We follow here closely his exposition and adopt his notation. Starting point is the following random field (a random field is – simply speaking – a mapping of a probability space in a d-dimensional space)

$$\mathbf{Z}(\mathbf{s}) : \mathbf{s} \in D \tag{3.1}$$

where $\mathbf{s} \in \Re^d$ is some location in d-dimensional Euclidian space and $\mathbf{Z}(\mathbf{s})$ is a random quantity at the location $\mathbf{s}_i$ that varies over the index set $D \subset \Re^s$. For most applications, D is assumed to be a fixed sub-set; however, treating D as random (sub-)set of $\Re^d$ allows for greater flexibility. By that, given the application, the randomness can be modeled either by varying $\mathbf{Z}$ or varying D from realization to realization. Therefore, D as a subset of $\Re^d$ is called a *spatial process*. In most applications either the $\mathbf{Z}$-process varies and the D-process is fixed or vice versa. Although models are possible in which both processes covary, usually independence or conditional independence is then assumed.

3.1 Three Classes of Spatial Data

Given the applications at hand commonly three classes of spatial data are differentiated within the field of spatial statistics (Cressie (1993), also compare for example Čižek et al. (2007) or Banerjee et al. (2004)):

- **geostatistical data** – sometimes called *spatially continuous data* (Čižek et al. 2007), *point-referenced data* or *geocoded data* (Banerjee et al. 2004),

- **point patterns** – also called *spatial point patterns* (Čižek et al. 2007), and
- **lattice data** – also called *areal data* (Čižek et al. 2007) or *aggregate data*.

In the following three sections we give a brief introduction to the modeling approaches for each kind of data and express them formally within the general spatial model of equation 3.1.

Geostatistical Data

Geostatistical data derive mostly from applications in mining and other geographical sciences. It those fields, spatial predictions are needed, such as the grade of an ore or the soil property for a whole area given measurements from a number of fixed locations. Two different kind of (spatial) effects have to be modeled. The large scale variation (first order effect) which describes a spatial overall trend of the data (e.g. a north-south trend) and the small scale variation (second order effect) which describes the spatial correlation – nearby measures are more related than distant measures.

It is assumed that D is a fixed subset of $\Re^d$ with d being positive and the spatial index $\mathbf{s}$ varies continuously through region D. We observe the realizations $Z(\mathbf{s_1}), \ldots, Z(\mathbf{s_n})$ from n prior fixed sampling locations. A crucial assumptions is that the covariance between $Cov(Z(\mathbf{s_i}), Z(\mathbf{s_j}))$ depends on the distance between these two locations. Different specification are used to

model the distance but a commonly used is the *exponential model* that assumes $Cov(Z(\mathbf{s_i}), (Z(\mathbf{s_j}))) \equiv C(d_{ij} = \sigma^2 \exp^{-\phi d_{ij}})$ for $i \neq j$. The variable d_{ij} is the distance between two locations, and σ^2 is the *partial sill* and ϕ is the *decay parameter*, both are positive. In the case of $i = j$ the distance is zero and the covariance becomes $Var(Z(\mathbf{s_i})) = \tau^2 + \sigma^2$, where τ^2 is called the *nugget effect* and the expression $\tau^2 + \sigma^2$ called the *sill*. A plot of the covariance is called a *covariogramm*. The term *kriging* refers to the making of predictions at site $\mathbf{s}_0$ which has not been observed given the known observations in $\mathbf{Z}(\mathbf{s})$.

Point Patterns

For point pattern processes D itself is random now and the interest lies what model drives the occurrence of an event $\mathbf{Z}(\mathbf{s})$ at a random location. Mostly $\mathbf{Z}(\mathbf{s})$ is set to equal 1. A good example for this is the occurrence of a disease or of a certain (binary) biological feature such as a the nest of certain species. If the occurrence itself carries some information, such as the epicenter of an earthquake with magnitude 1 or 2, the process is called then a *marked point process*.

A general question of interest for point patterns is whether clusters can be identified or if the occurrences of the events are completely at random. Graphical approaches have proved to be futile as humans tend to identify clusters even in pattern determined completely by chance. Furthermore, clusters can also occur in random processes, the question is now to what extend is the identified cluster beyond chance?

Starting point for the description of uniformity is *homogeneous Poisson process*: the expected number of occurrences in a given region A is $\lambda|A|$, where λ is the intensity parameter and $|A|$ is the area of A. A well established statistic to test for clustering is Ripley's K function:

$$K(d) = \frac{1}{\lambda} \mathrm{E}[\text{number of points within } d \text{ of an arbitrary point}] \tag{3.2}$$

The parameter λ can be estimated as the mean number of points per unit area. For point processes that have no spatial dependencies the values for K take the form $K(d) = \frac{\pi}{d^2}$. The number of points within a distance d should increase proportional to the area of a circle with the radius d. For a cluster we would observe $K(d) > \pi d^2$ and for some randomly spaced pattern $K(d) < \pi d^2$. Inference can now be made by comparing the estimate of K with a theoretical quantity. The standard estimator for K is: $\hat{K} = n^{-2}|A| \sum \sum_{i \neq j} p_{ij}^{-1} I_d(d_{ij})$, where n is the number of points in area |A|, d_{ij} the distance between points i and j, p_{ij} the proportion of the circle with center i and passing through j that lies within A, and $I_d(d_{ij})$ is an Indicators variable that equals 1 if $d_{ij} < 0$ and 0 otherwise.

Lattice Data

Starting point for areal data is a random process $\mathbf{Z}(\mathbf{A_i}) : A_i \in (A_1, \ldots, A_n)$ where A_i are regions that are elements of D that is referred to as a (spatial)

lattice[1]. Furthermore, $(A_1, \ldots, A_n)$ is a partition such that $A_1 \cup A_2 \cup \ldots \cup A_n = D$ and $A_i \cap A_j = 0$ for $i \neq j$. The spatial structure has to be modeled explicitly via a neighborhood information matrix called $\mathbf{W}$ consisting of $N \times N$ elements indicator the spatial relationship between region A_i and A_j at weight matrix element w_{ij}.

The reason for analyzing areal data are usually that some data can only be measured at some aggregate level although the actual process generating such data is at the individual level. For example, unemployment figures are an aggregate of the binary outcome employed/unemployed which – in theory – could be measured at the precise location of the individual in space. Another reason is often that this data is not publicly available due to anonymity concerns and only aggregate data is released, this is often true for disease and mortality data. The boundaries used are often somewhat arbitrarily as these are some administrative structure such as zip-codes or municipalities. Often these boundaries have no influence on the underlying stochastic process. However, also the opposite can be true, such as when the area in question has the authority to alter policy and by that influences the process in question, e.g. different local employment policies.

[1]The term lattice data seems somewhat misleading as this kind of data not only includes regularly spaced areal data, such as from agricultural trials or pixeled images, but also irregular formed areas, such as arising from administrative boundaries. But the term is well established in the spatial statistics field and encompasses all kind of areal data. Hence, the term lattice data and areal data are used interchangeably.

3.2 Neighborhood Weights

For the spatial analysis of areal data the definition of the weight matrix W – sometimes called *proximity matrix* or *connectivity* matrix – is crucial as it carries the spatial information for lattice data (for the following compare Haining (2005)) . The matrix $\mathbf{W}$ has the dimension $N \times N$ where N is the number of regions in the data set and consists of the elements w_{ij} that indicates the spatial relation between region i and region j. Conventionally, the values for the diagonal are set to zero.

Obviously several ways exists to specify the spatial relations between regions as indicated by $\mathbf{W}$. However, no rule exist for the optimal choice. In fact, the functional specification of the distance matrix is an open and controversial question within the field of spatial statistics. If no or little theocratical grounds exist in defining a weight matrices, it is often suggested to try several different definitions to gauge the change in the estimations. Some suggest, that the weight matrix should be used, that yields the highest spatial correlation coefficient λ in the spatial regression (see section 3.5), other argue the weight matrix should be used that provides the best overall model fit (Chi and Zhu 2008).

In the following we introduce the most common approaches used in the literature. Furthermore, we assume that we deal with lattice data (not in the most strictest sense), therefore every region has at least one border to another region and every region can be reached from another region by connecting regions (*contiguity*). Otherwise stated, we exclude islands (careless if the consist

of one or of several regions).

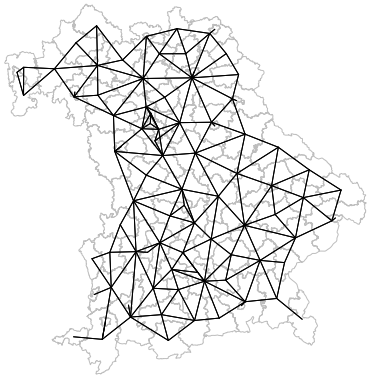

Figure 3.1: Example for a single contiguity neighborhood structure based on *Bayern*

Figure 3.2: Example for a 4-nn neighborhood structure based on *Bayern*

Contiguity Based Spatial Weights

The most simple weight matrix is the *simple contiguity* weight matrix. If the region i and the region j share a border they are considered neighbors. The

value 1 indicates the status of being direct neighbors and zero otherwise. This can be formally expressed as:

$$w_{ij} = \begin{cases} 1 & \text{if regions } i \text{ and } j \text{ share a border} \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

The simple contiguity weight is the most used weight in spatial analysis as it is rather easy to calculate and most easy justifiable on theoretical grounds. But also *higher order* contiguity weights are used. In the case of a contiguity matrix of order 2, a region j that does share a border with region i, w_{ij} is set to 2. For a region k that does not share a border with region i but with a region j (that shares a border with i), w_{ik} is set to 1. This can be, in principal, done for an arbitrary order, where the highest value for w_{ij} shows direct neighborhood and a lower value shows a larger distance (measured in neighborhood steps). In Figure 3.1 an example is given for a a neighborhood structure based on the single contiguity criterion. Neighborhood structure maps for data with many regions or with higher order contiguity are often not very intelligible.

Distance Band Spatial Weights

In the case of *distance bands* based spatial weights, a cut-off value c is chosen. Every region that is within this distance of the i-th region is considered a neighbor. The distance is usually measured from the centroid of the respective

regions.

$$w_{ij} = \begin{cases} 1 & \text{if } d_{ij} < c \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

Of course, higher order spatial weights are conceivable using distance bands. Several cut-off values can be specified and assigned in the following fashion:

$$w_{ij} = \begin{cases} p & \text{if } d_{ij} < c_1 \\ p-1 & \text{if } c_1 \leq d_{ij} < c_2 \\ \vdots & \\ 1 & \text{if } c_{p-1} \leq d_{ij} < c_p \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

In the case of irregular sized or shaped regions (which is often the case) a reliable and comparable measurement of distance is not easily found.

K-Nearest Neighbor Spatial Weights

In the case of the *K-Nearest Neighbor* approach simply the k closest regions of region i are considered neighbors. The number chosen for k should be a rather small integer. Sometimes $k = 1$ is selected to mark only the closest neighbor. In Figure 3.2 an example is given for a 4-nn neighborhood structure for the region of *Bayern*. It becomes evident that it carries a different information than just a single contiguity neighborhood structure (see Figure 3.1), it rather neatly divides the region of *Bayern* in a northern and a southern part.[2]

[2] A different question is whether this is theoretical meaningful or not.

$$w_{ij} = \begin{cases} 1 & \text{if rank of } d_{ij} \leq k \\ 0 & \text{otherwise} \end{cases} \tag{3.6}$$

General Distance Weights

Other ways to construct the weight matrix, that move beyond the neighborhood approach, have been suggested. These are based on some distance measure. The most simple is given by the *inverse distance weights* with $w_{ij} = d_{ij}^{-\delta}$. The distance d can be calculated by different metrics such as Euclidean, City-Block-Distance, or any other that makes theoretical sense such as traveling cost.[3] Another possibility lies in employing the *exponential function of distance*: $w_{ij} = \exp^{d_{ij}^{-\delta}}$. In both cases the value of a weight decreases with increasing distance. The steepness of this decaying function can be controlled with parameter $\delta > 0$.

The *common border* function is given by $w_{ij} = (\frac{l_{ij}}{l_i})^{\tau}$, where l_{ij} is the length of the common border between region i and region j, and l_i is the total length of the border of region i. The weighting matrix is only non-zero when a border is shared and this value diminishes when the shared border is small and vice versa. The parameter τ influences the steepness of this decrease. This kind of weights are in particular interesting when exposures has to be modeled, such as from an environmental hazard.

[3]For distance measures using large areas, one has to take into account the curvature of the earth. The shortest distance (*geodesic*) between two points is calculated by $D = R\phi$, where R is the radius of the earth and ϕ is an angle between two point at the center of the earth. Hence, the distance becomes the length of the arc of a circle with radius R (Banerjee et al. 2004).

In some ways, the use of general distance based measures is the use of geostatistical methods (Wall 2004) as outlined in section 3.1. Instead of having geo-referenced points, the centroid of the region is taken as a single geo-referenced point and all observations in the regions are summed up and assigned to that centroid. The thrust of criticisms against using this kind of weights, is that the choice of the centroid is as the geo-referenced point is somewhat arbitrary. This becomes in particular clear if the size of the regions vary dramatically.

3.3 Choropleth and Probability Maps

The aim of mapping a statistical variable is to visualize the spatial distribution of the data. The observer might want to assess if there is a overall (spatial) trend in the data or if there are clusters of particular low or high values. Mapping of areal data faces two basic choices: Either mapping polygons or just *centroids* – the points where the center of each polygon lies. The latter option (sometimes called *cartogramms*), is an interesting option as one can alter the size of the centroid to show the magnitude of the variable in question. However, the picture will be rather abstract and one is losing the additional information a map provides. To depict the values of a continuous variable on a polygon map, the variable values are divided into several categories and every class is assigned a color to represent a particular value (hence a *choropleth* map). The number of classes is predefined and the cut-off rule can follow some

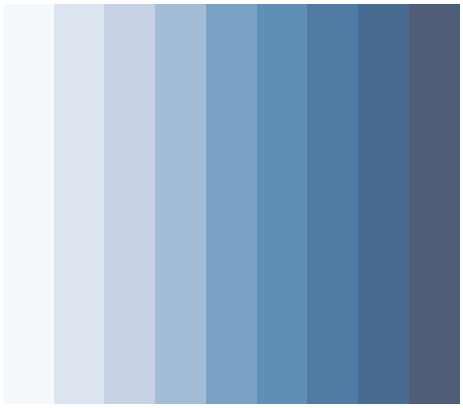

Figure 3.3: Sequential palette of blue colors used for thematic mapping (dark colors for high values and light colors for low values)

explicit criteria such as quartiles, quintiles, standard deviations, or certain algorithm following some optimization rule – but usually categories of equal length are defined. The idea is to show a continuum of colors that depict low to high values. A crucial question is the choice of the color scheme for mapping values.

Certainly, using a lot of colors has an aesthetical appeal and is popular among consumers of maps. But cognitive based research showed in a string of studies, that humans have problems with colored maps (compare Lawson (2001) p. 33 with further references). Although humans do prefer colored maps and they *believe* they can infer more information from such a map (e.g. recognitions of trends or clusters) they actually do worse than with a monochromatic map. Hence, *intensity* of a *single* color should be used. This view is also supported by Tufte (1999) who points out that there is no universally accepted hierarchy of colors. Tufte, furthermore, suggests that a gray scale should al-

ways be used for the sake of clarity, we refrain, however, from his – rather rigid – recommendation and use a scale of blue colors (see Figure 3.3; dark colors signify high values and light colors low values).

A useful extension of choropleth maps are so called *probability* maps (Cressie 1993) . The cut-off of the classes is chosen to show the probability of the occurrence of the observed value. For example, an observation is within the .1 percentile, the 1st percentile, or the 5th percentile of the empirical distribution. This kind of cut-off rule is very helpful in identifying unusual high or low valued observations and deciding whether a significant hot spot exists or not. However, attentions has to be paid whether the empirical distribution complies with the assumed distribution to calculate the probability. Furthermore, in a large data set a few very high or low observations are part of the natural variability of the data and should not be to lightly judged as hot spots.

3.4 Measuring Spatial Autocorrelation

When searching for *spatial autocorrelation*[4] within lattice data measures are divide into *global* and *local* measures of spatial autocorrelation. The former assess whether the data as a *whole* exhibit spatial autocorrelation, tested versus the assumption of no spatial randomness. The latter tries to identify *particular* observations that are significantly autocorrelated with neighboring

[4]Depending on the literature also called *spatial dependence* or *spatial interaction* (Chi and Zhu 2008).

observations (Darmofal 2007). Here we introduce *Moran's I* as a standard measure for global spatial autocorrelation:

$$I_{Moran} = \frac{N}{S} \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2} \tag{3.7}$$

where N is the number of regions, $S = \sum_i \sum_j w_{ij}$ the sum of the weights as a scaling factor, w_{ij} is the denoted element in the weight matrix $\mathbf{W}$, y_i and y_j give the values at the respective regions, and $\bar{y}$ is the overall mean.

Morans I differs in two respects from a standard correlation coefficient. First, in the case of no correlation the value does not become zero but $\frac{-1}{N-1}$. However, it is a function of the sample size and approaches zero rapidly when N increases. Second, the interval of $[-1, 1]$ is not fully supported, the resulting range is slightly narrower. Inference about the significance of Moran's I can take two form, either using the normal distribution or a Monte Carlo approach. The former, however, is not recommended (Banerjee et al. 2004). The other approach is a Monte Carlo based permutation approach. As the distribution of I is invariant to permutations of Y_i under the null hypothesis of spatial independence. However, this would require to calculate $N!$ permutations. To avoid this burdensome approach, a Monte Carlo sample of n permutations is drawn, usually 1000, including the observed data. Now the the the I of the observed data will be ranked withing a the distribution of the other $n - 1$ randomly derived $I's$ to obtain an empirical p-value.

3.5 Spatial Error Model

The starting point for the specification of a regression model[5] for spatial areal data is the standard regression model – we follow exposition taken in (Anselin 2006), (LeSage 1999), or (Anselin 1992). The model can be expressed as:

$$y_i = \sum_k x_{ik} \beta_k + \epsilon_i \tag{3.8}$$

where y_i is now an observation for the dependent variable in region i and x_{ik} denotes observations for the k-th independent variable for a given region. Usually, it includes a constant. The variable β are matching regression coefficient and ϵ_i is a random error term. In the classic regression model the error term is assumed to be identically and independently distributed (*i.i.d.*). Hence, the error term is normally distributed with $\epsilon \sim N(0, \sigma^2)$.

The spatial dependence of the observations, that violates the independence assumption, can now be accounted for via the error term. This is the so called *spatial error models* and it accounts for spills effect across neighboring regions that cannot be modeled because of lack of an appropriate covariate. This lack might be due to incomplete theory or inability to measure.

Modeling the spatial dependency through the error term is a special case of a non-spherical error covariance matrix in which the off-diagonal elements are correlated $\mathrm{E}[\epsilon_i \epsilon_j] \neq 0$. The values of the off-diagonal elements depend on

[5]The rationale here is to start from a classical regression model and accommodate the spatial relationships via the error term. This leads to a smaller class of models than the approach taken by Cressie (1993), but allows a more intuitive understanding.

the spatial ordering of observations. Closer regions have a higher correlation than more distant ones. To estimate the covariance matrix some restrictions must be imposed (as without the restriction $N \times (N-1)/2$ parameters would be needed). A way to impose this structure is to specify a spatial process for the random disturbances. A common choice is to model ϵ as a *spatial autoregressive process*:

$$\epsilon_i = \lambda \sum_j w_{ij}\epsilon_j + u_i \tag{3.9}$$

In equation 3.9 the variable λ is the autoregressive parameter and w_{ij} is an element of the weight matrix $\mathbf{W}$ as introduced in section 3.2. The random error term u_i is assumed to be normally distributed. This equation to can be expressed in matrix notation and then rearranged to yield:

$$\epsilon = \lambda \mathbf{W}\epsilon + \mathbf{u} = (\mathbf{I} - \lambda \mathbf{W})^{-1}\mathbf{u} \tag{3.10}$$

with $\mathrm{E}[\mathbf{uu^{th}}] = \sigma^2\mathbf{I}$. The variance-covariance matrix of the error term follows:

$$\mathrm{E}[\epsilon\epsilon^\top] = \sigma^2(\mathbf{I} - \lambda \mathbf{W})^{-1}(\mathbf{I} - \lambda \mathbf{W}) \tag{3.11}$$

The role of the matrix $\mathbf{W}$ is certainly crucial now. It is important to note that it has full rank as we defined it only for regions with at least one neighbor. The spatial error model can now be compactly expressed as:

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\beta + \epsilon \\ \epsilon &= \lambda \mathbf{W}\epsilon + \mathbf{u} \\ \mathbf{u} &\sim N(0, \sigma^2) \end{aligned} \tag{3.12}$$

An inherent problem with the spatial error model is that it prone to heteroscedasticity due to spatial heterogeneity. This can often have a simple cause, such as different population sizes in the respective regions, but also due to spatial non-stationarity of the data. Hence, often a spatial trend variable is constructed to account for heteroscedasticity. Whether the used model accounted sufficiently for heteroscedasticity can be tested for by using a spatial version Breusch-Pagan test that takes λ into account.

The outlined spatial error model is sometimes called a *simultaneous autoregressive model* (SAR) (Cressie 1993) as it assumes a simultaneous spatial process. The spatial autoregressive process does not necessarily has to be modeled through the error term. Sometimes this is done through a spatial lag model: $Y = \rho\mathbf{W}\mathbf{Y} + \mathbf{X}\beta + \mathbf{u}$, where the spatially lagged dependent variable is included in the right hand side and has a substantial interpretation. Another very often used variant is the so called *conditionally autoregressive model* in which the spatial effect on region i is only conditionally on the neighboring regions and not simultaneous across the lattice. Whereas the *spatial error model* explain the relations among response variables at all locations on the lattice simultaneously and the spatial effect is considered to be endogenous, the CAR

models specify the distribution of the response variable at one location by conditioning on the values of its neighbors in the neighborhood and the spatial effect of the neighbors is considered to be exogenous.

Estimation

When the spatial parameter λ is known, then the estimation is straightforward using well established OLS. But this is seldom the case and λ has to be estimated simultaneously from the data. Ord (1975) suggested an iterative procedure based on time series models:

1. Compute the OLS residuals ($\tilde{\epsilon}$) from $\mathbf{y} = \mathbf{X}\beta + \epsilon$
2. Estimate λ from $\tilde{\epsilon} = \lambda\mathbf{W}\tilde{\epsilon}$ using ML (see below); the estimate is called $\tilde{\lambda}$
3. Construct the new variables $\tilde{z} = (I - \tilde{\lambda}\mathbf{W})\mathbf{y}$ and $\tilde{\mathbf{X}} = (\mathbf{I} - \tilde{\lambda}\mathbf{W})\mathbf{X}$
4. Apply OLS for $\tilde{\mathbf{z}}$ on $\tilde{\mathbf{X}}$ to get a new estimate for $\tilde{\beta}$
5. Construct the new residuals $\tilde{\epsilon} = \mathbf{y} = \mathbf{X}\tilde{\beta}$; repeat until convergence is achieved

In step 2 a estimate for λ is needed. This is done using a maximum likelihood approach. The main problem with a ML-approach is that the Jacobian determinant $\ln|\mathbf{I} - \lambda\mathbf{W}|$ must be evaluated for every iteration. The standard solution is to decompose the log Jacobian into $\sum_i \ln(1 - \lambda\omega_i)$, where ω_i are the

eigenvalues[6] of the weight matrix $\mathbf{W}$. Hence, the log-likelihood can be shown to be (Anselin 2003):

$$\ln L = \sum_i \ln(1 - \lambda\omega_i) - \frac{N}{2}\ln 2\pi - \frac{N}{2}\ln\sigma^2 - \frac{(\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{I} - \lambda\mathbf{W})^\top(\mathbf{y} - \mathbf{X}\beta)}{2\sigma^2}$$

[6]Do not mistake it with w_{ij}, an element of the weight matrix.

Chapter 4

Data and Analysis

4.1 Data

Map and Neighborhood Weights

The map used for the analysis is an ESRI-shape[1] file published by the Bundesamt für Kartographie (2004) and shows the administrative boundaries for all municipalities of the Federal Republic of Germany. For the purpose of the analysis the file had to be altered manually. First, to avoid the island effect, the county of Rügen had to be discarded (compare page 26). Second, for official mapping purposes the Bodensee-lake is drawn as two separate regions, both of them had to be discarded as well. Hence, the map contains 438 observations.

[1] ESRI-shape files are a commercial standard introduced by a company of the same and form a quasi-standard for digital maps.

As pointed out on page 26 the modeling choices for the neighborhood structure are plentiful and somewhat arbitrary yet not without effect. We propose to use two different contiguity weights that are established in this line of research questions (Chi and Zhu 2008), namely: single and double contiguity.[2] The double contiguity weight (weight matrix of the order two) is chosen to account for the fact that some regions only have one neighbor under the single contiguity criterion. This is mostly due to cities that are surrounded by a single region.

Variables

The data were taken from *Statistik Regional*, a DVD compiled jointly by the Bundesamt for Statistik and the statistical offices of the *Länder* (Bundesamt für Statistik 2007). The 2007 edition was used, where the most recent mortality data was from the year 2004. Hence, all data used in this analysis is from 2004. The dependent variable is the standard mortality ratio for all cause mortality between the age of 20 to 49. In face of the outlined demographic theory, the available data, the need to conduct parsimonious modeling, and to avoid the pitfalls of data mining we propose four regressors: income, number of hospital beds, migration rate, and a spatial trend.

[2]We refrain from the use of distance based measure as the calculation of distances for a shape file covering such a large and irregular shaped area is non-trivial task and beyond the scope of this thesis.

Dependent Variable The dependent variable in this analysis is the standard mortality ratio for all causes of death for the age groups between 20 and 49 years of age for the year 2004. The choice for this restricted age group is partly caused by the problem of getting reliable and stable rates for certain age groups. Moreover, this age group has the highest propensity to migrate, making it most appropriate to study the healthy migrant hypothesis. The calculations are only based on German citizens (hence, no foreigners with resident permits) and is not differentiate by sex. The ratio was calculated using the values for the national population of Germans as the standard. The original data had age groups with five year intervals.

In Figure 4.1 and 4.3 a boxplot and a histogram, respectively, of the SMR is shown. It has a unimodal distribution, however, it contains some very large observations (compared to the bulk of the data). The question arises whether it is useful to take the logarithm of the SMR. An overview of some descriptive statistics is given in Table 4.1. The comparison between the measures for skew and kurtosis show that the log of SMR is less skewed and much close to the values expected for a normal distribution. This is also supported by a comparison of the respective QQ-plots in Figures 4.5 and 4.6. The logarithm of SMR is less skewed and has only one observation outside the bulk of the data (see Figure 4.2). Hence, we are going to use the log of SMR as the dependent variable and treat it as sufficiently normal distributed.[3]

[3]Strictly speaking the numbers deaths are count data and imply a poison regression approach. However, considering the large number of counts a approximation by the normal distribution is in principle reasonable and well established in the literature (see for example (1981) with further references).

Table 4.1: Descriptive statistics of SMR and log of SMR

var	n	mean	sd	median	min	max	skew	kurtosis
SMR	438	1.03	0.23	1.00	0.52	2.09	0.69	0.91
log of SMR	438	0.01	0.22	0.00	−0.65	0.74	0.05	−0.15

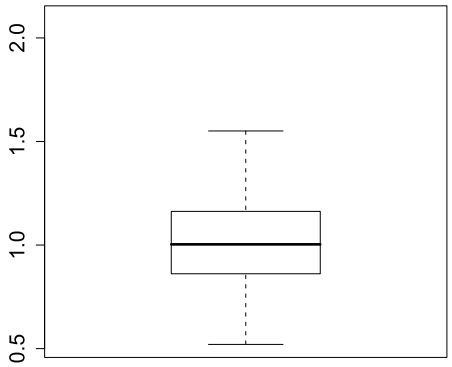

Figure 4.1: Boxplot of SMR

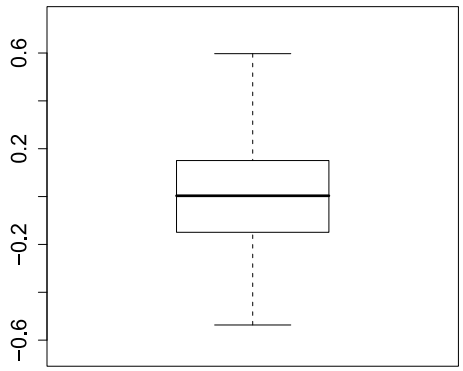

Figure 4.2: Boxplot of log SMR

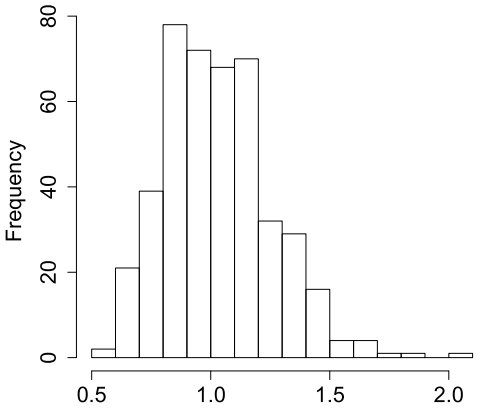

Figure 4.3: Histogram of SMR

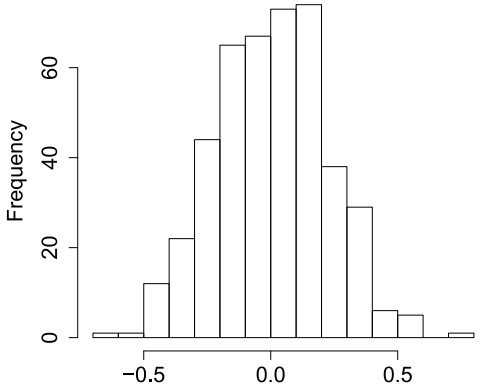

Figure 4.4: Histogram of log SMR

Independent Variables The first variable (INC) is income measured by the per-capita income in the respective region. To avoid the problems of a skewed distribution this variable was transformed using the natural logarithm.

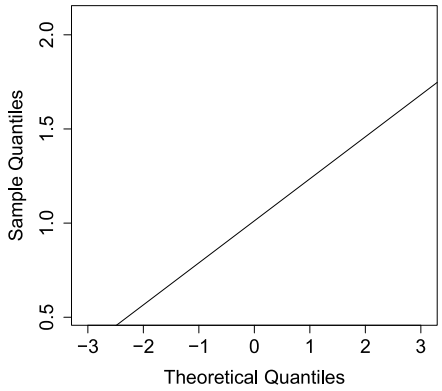

Figure 4.5: QQ-plot of SMR

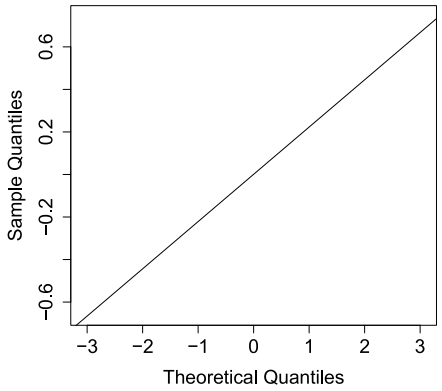

Figure 4.6: QQ-plot of log of SMR

The second variable (MIG) is the migration rate and was derived by subtracting the number of all departures[4] by the number of all arrivals (of German citizens). To make this figure comparable, it was divided by the total population. For ease of interpretation is was then multiplied by 1,000; hence yielding the net-out-migration rate per 1,000 residents. Unfortunately, migration data was not available by age groups. This introduces a small measurement error, but one has to keep in mind that migration is a phenomenon which affects the age groups between 20 and 50 years of age the most. The variable can be more easily considered as a measure of the direction and strength of migration flows for the respective region.

The third variable (BED) is the medical infrastructure operationalized by using the number of hospital beds in the municipality. The variable was divided by the total population of the region and the multiplied by 1,000 yielding the number of hospital beds per 1,000 citizens. The variable was highly skewed,

[4]Defined as registering in a municipality outside of the respective region.

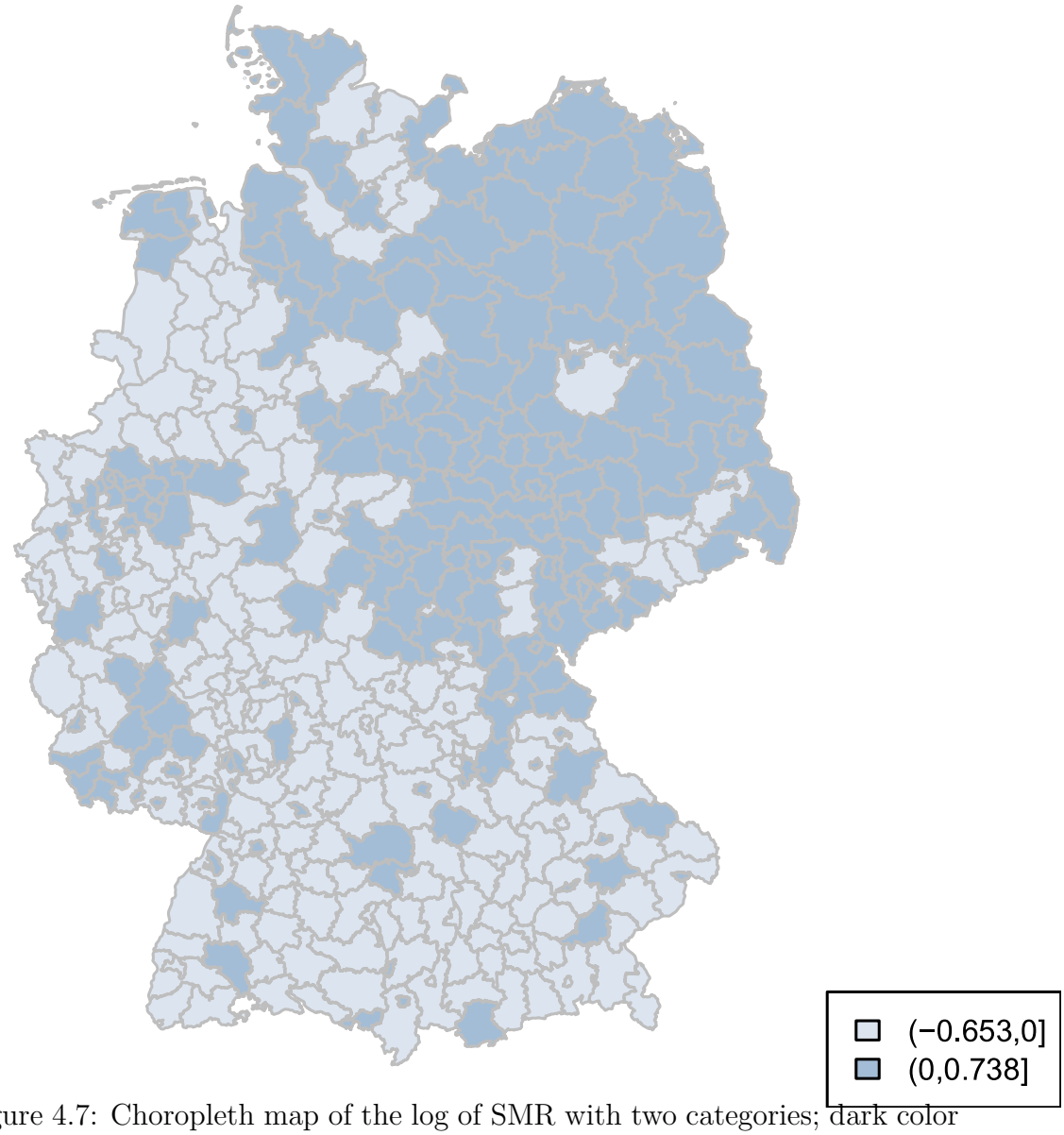

Figure 4.7: Choropleth map of the log of SMR with two categories; dark color indicates a mortality higher than the country average and light color indicates a mortality lower than the country average

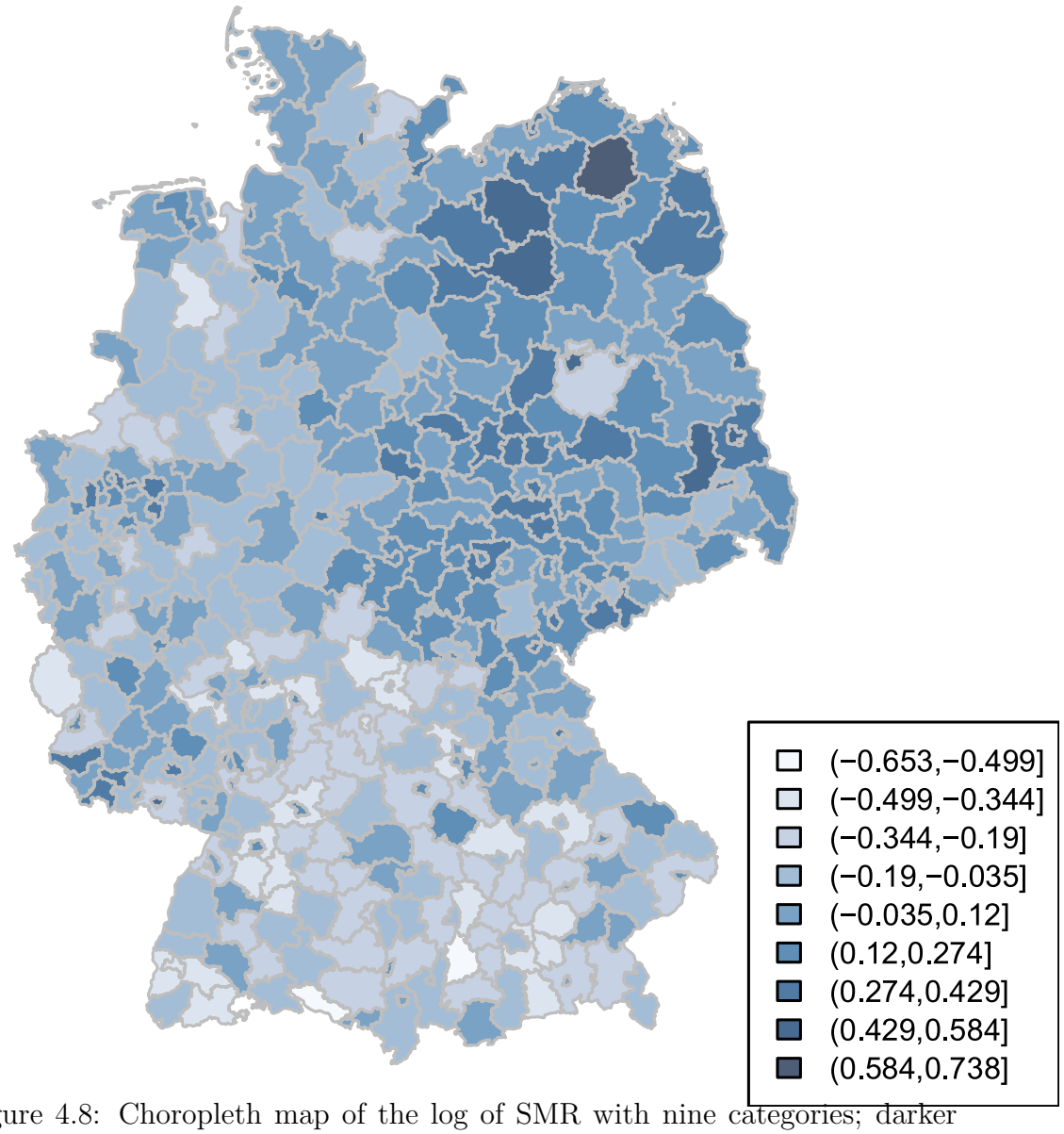

Figure 4.8: Choropleth map of the log of SMR with nine categories; darker colors indicate a higher mortality (see legend)

therefore it was transformed using the natural logarithm. To deal with the observations that had a zero, a constant with the value 0.5 was added to all observations before the transformation.

Finally the last variable is the spatial trend. This variable is of greater importance. Not only as it has a substantially interesting interpretation but also because it used to mitigate spatial heteroscedasticity. A visual inspection of Figure 4.7 indicates an East-West trend. However, Figure 4.8 rather indicates a North-South or even a North/East to South /West trend. We are going to use trends for all three meaningful possibilities (N-to-S-trend, E-to-W-trend, and NE-to-SW-trend) and compare the outcomes. For the ease of interpretation we normalized the trend variables by subtracting the mean and dividing it by the standard deviation. Hence a value of zero indicates a central position on the map.

Table 4.2: Descriptive statistics of the independent variables

variable name	n	mean	sd	median	min	max
INC	438	9.73	0.13	9.74	9.47	10.23
BED per 1000	438	2.73	0.64	2.76	−0.69	4.19
MIG 1000	438	0.74	11.54	−1.07	−35.48	52.38
NE-to-SW-trend	438	0.00	1.00	−0.17	−2.04	2.37
E-to-W-trend	438	0.00	1.00	0.01	−1.9	2.31
N-to-S-trend	438	0.00	1.00	0.02	−1.82	2.32

4.2 Analysis

Initial data analysis

The initial data analysis consists of two steps. First, we need to establish whether the dependent variable exhibits a spatial dependency structure. Second, we test for bivariate associations of the independent variables with the dependent variable.

Testing for Spatial Correlation

The test for spatial autocorrelation is done using Moran's I (see page 33). It is important to remember, that it is not recommended to use common test statistics but conduct a Monte Carlo based permutation test. The results for the two different weight matrices used are given in Figure 4.9 and 4.10, respectively; clearly both coefficients are significant and possess a sizable magnitude ($I_{MORAN} = .42$ for the single contiguity and $I_{MORAN} = .41$ for the double contiguity weight matrix). Hence, it is clear that the spatial distribution of the dependent variable differs systematical and is not random.

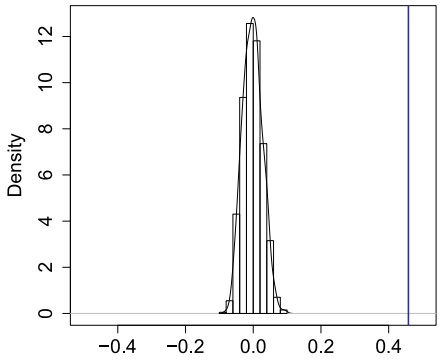

Figure 4.9: Permutation test for Moran's I for log of SMR with single contiguity weights (horizontal line indicates location of empirical I compared to reference distribution)

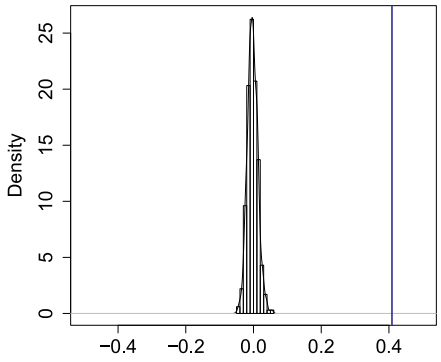

Figure 4.10: Permutation test for Moran's I for log of SMR with double contiguity weights (horizontal line indicates location of empirical I compared to reference distribution)

Bivariate Association

The scatter plots of the independent variables with the dependent variable indicate a linear relationship.[5]

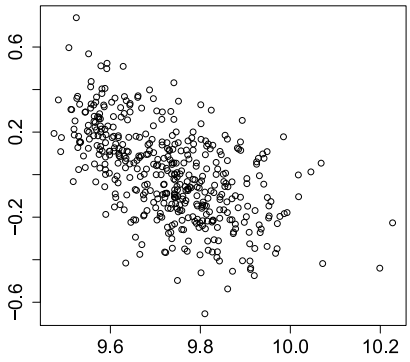

Figure 4.11: Scatter plot of INC vs. log of SMR

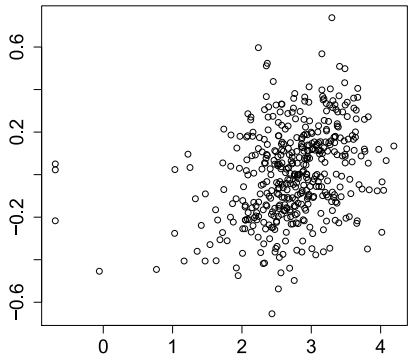

Figure 4.12: Scatter plot of HOS vs. log of SMR

[5]Doubts about a linear relationship could be cast for the East-West-trend variable (see Figure 4.16)

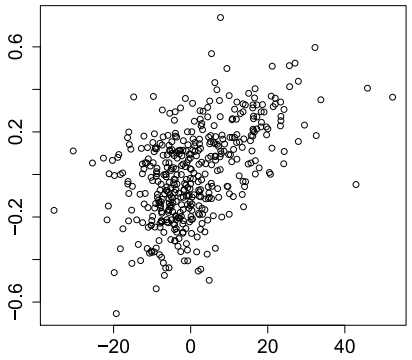

Figure 4.13: Scatter plot of MIG vs. log of SMR

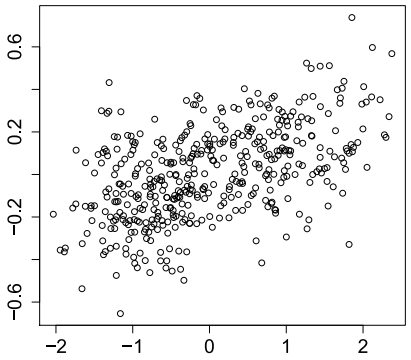

Figure 4.14: Scatter plot of NE-to-SW-trend vs. log of SMR

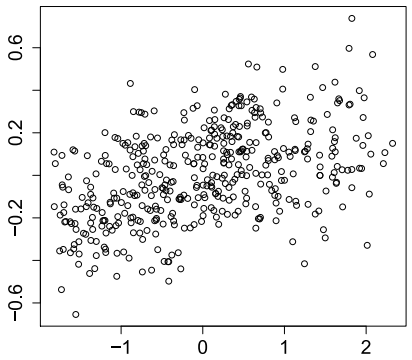

Figure 4.15: Scatter plot of N-to-S-trend vs. log of SMR

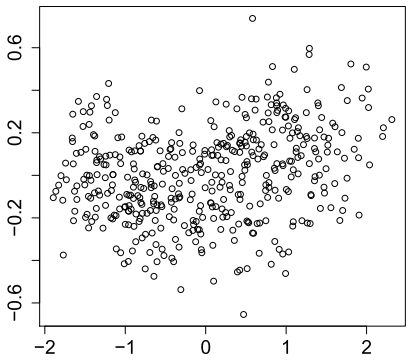

Figure 4.16: Scatter plot of E-to-W-trend vs. log of SMR

Spatial Error Regression Model

The chosen regression specification is a *spatial error model* with two different weight matrices and three different trends. Hence, six regressions were run. An overview of the results is given Table 4.3. All six models correct sufficiently for spatial autocorrelation. Moran's I is in all six models practically

zero[6] and non-significant (output omitted). Furthermore, in all models is the spatial version of the Breusch-Pagan-test not significant, hence we do not reject the assumption of homoscedasticity. For all residuals we cannot reject the assumption of a normal distribution as shown by the non-significance of the Jarque-Bera-test. Hence, all regression models are able to sufficiently model the spatial structure of the data.

The best overall model fit using the AIC-criterion has model two. It also has the second highest λ. Comparing the different operationalizations of the trend it is interesting to note, that the model with the highest AIC also has the largest coefficient for the trend variable. It is the Nort/East-to-South /West-trend with a coefficient of .062. However, this value is not significant different from the coefficient for the East-to-West-trend. The North-to-South-trend is not significantly different from zero. The other covariates are all significant at the 99% level. The respective regression specification do not differ much by the spatial weight. And only for the the regression models using a East-to-West trend, the double contiguity matrix has a higher AIC than the single contiguity matrix. Hence, a single contiguity matrix seems sufficient to model the spatial structure of the data.

For all six regression specification the coefficients have a very similar magnitude. This increases our trust in the stability of the estimation results. In the discussion we focus on model two as it has the best overall model fit. Income has a negative effect on the SMR. An increase in income leads to a decrease

[6]Remember that Moran's I in case of no correlation not zero but $\frac{-1}{N-1}$ which is in this application -.002.

in mortality. This is an expected connection. For hospital beds, however, the effect is positive – implying that an increase in medical infrastructure leads to increased mortality. This is certainly a questionable relationship. It rather seems, that regions that have a high mortality tend to have more hospital beds, by that refuting the hypothesis of medical underspending. The variable MIG shows that regions that out-migration increases the SMR of a regions. This implies that a healthy migrant effect exists – it is the less frail who leave their regions. The coefficient for the trend variable – in model two we specified a North/East-to-Sout/West-trend – implies that regions to the North-East have a higher SMR than regions to the South-West. The log-log specification of the income variable allows for an convenient interpretation of the magnitude Wooldridge (2005). An increase in the per capita income by 10 % would lead to a decrease of the SMR by 4.7 %. Assuming the national mortality is unaltered, this implies also 4.7% less observed deaths. For the log-level specification used with the migration rate variable we can infer, that an increase by one unit leads to an increase in the SMR by less than a half percent. The trend variable, which has a log-level specification as well, implies that moving one unit to the North-East the SMR increases by 6.2%. Considering that the trend variable has a range of about four (compare Table 4.1), the spatial trend variable explains a large amount of the variation of the mortality.

Table 4.3: Regression results

Model No.	1	2	3	4	5	6
Specification of TREND	NE to SW	NE to SW	E-to-W	E-to-W	N-to-S	N-to-S
Weight Matrix	double	single	double	single	double	single
CONSTANT	3.762	4.317	4.616	5.920	4.002	4.834
	(0.843)	(0.832)	(0.822)	(0.803)	(0.831)	(0.799)
INC	-0.411	-0.470	-0.494	-0.633	-0.435	-0.523
	(0.087)	(0.086)	(0.085)	(0.083)	(0.085)	(0.082)
BED	0.095	0.099	0.086	0.094	0.094	0.098
	(0.011	(0.011)	(0.012)	(0.011)	(0.011)	(0.011)
MIG	0.004	0.004	0.004	0.004	0.004	0.004
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
TREND	0.060	0.062	0.020	0.017	0.055	0.055
	(0.014)	(0.012)	(0.015)	(0.012)	(0.013)	(0.011)
λ	0.031	0.065	0.038	0.079	0.031	0.063
AIC	-423.04	-425.90	-408.24	-401.37	-420.92	-422.68
I_{MORAN}	-0.058	-0.013	-0.058	-0.022	-0.059	-0.009
BP-Test	5.888	6.0905	3.9059	1.2485	2.8719	2.9522
p-value	0.2077	0.1925	0.4189	0.87	0.5795	0.5659
JB-Test	0.3533	0.113	0.165	0.4324	0.348	0.060
p-value	0.838	0.944	0.920	0.805	0.840	0.907

Note: standard errors for regression coefficients are in parenthesis;
λ is the spatial autoregressive parameter;
BP-Test is the test statistics for the spatial version of the Breusch-Pagan-test and follows a χ^2-distribution with 4 df's (p-value in parenthesis);
JB-Test is the test statistic Jarque-Bera-test for normality and follows a χ^2-distribution with 2 df's.

Identifying Hot Spots

The second question is whether we can identify unusually high or low occurrences of mortality. Several methods exist (see Gómez Rubio et al. (2003) for some examples). Here we take an intuitive approach using the residuals from the spatial regression models. The underlying rationale is that the regression model does control for the spatial nature of the process – here we profit from the fact that the residuals are already corrected sufficiently for their spatial dependence (as seen by the non-significance of Moran's I) – and should includes all relevant variables explaining mortality. A region that has an unusual large absolute residual cannot be explained by the regression model and is hence an unusual hot spot. Clearly, it is more than doubtful that the regression model truly incorporates all relevant explanatory variables that explain regional mortality differentials. But regions that are not captured by the model certainly warrant a closer look to identify possible sources for the unusual high or low mortality. For the analysis we use the residuals from model two that had the best overall model fit. We compare the regions with a choropleth map of the log of SMR. The latter does not account for the spatial nature of the process. Both, the log of SMR and the residuals of model two are normally distributed. An overview is given in Table 4.4.

When just the log of SMR is taken into account, only two regions have a mortality that places them in the 1st percentile of the distribution and six regions have a mortality that places them in the top percentile. Comparing this with the findings implied by the residuals, the picture changes somewhat. Now

Table 4.4: Regions that have an unusual high or low mortality (the ID numbers are given, compare section B)

Percentile	log of SMR	Residuals of Model two
<.01	416 434	32 260 279 416 434
<.05	32 90 271 273 279 291 293 298 317 335 351 362 368 372 381 394 395 406 411 417 422 437	53 77 90 99 176 225 255 271 288 315 343 381 406 424
>.95	7 12 23 57 78 86 118 1 149 153 173 182 203 261 361	20 24 30 41 42 55 59 83 86 87 147 169 217 282 294 323 324 329 344 363 377 407 418 428
>.99	55 59 62 83 147	62 361

five regions are placed in the 1st percentile and only two regions in the top percentile. Hence, when controlling for the variables used in the regression, only two regions have an unusual high mortality compared to the national average. But five regions have now an unusually low mortality. Most notably is region 260. When just the log of SMR is used, this region not in the bottom five percent of all observations. However, when using the residuals it is placed in the first percentile, implying that it has – controlling for explanatory variables – a very low mortality. Certainly, those regions warrant a closer look to assess what might be the reasons for their significantly higher or lower mortality (an overview of the regions is given in the following maps). This comes, however, with a caveat; we have 438 observations so we do expect about 4 or 5 observations in the top and the bottom percentile, respectively. So those hot spots just might be a result of the natural variation of the data.

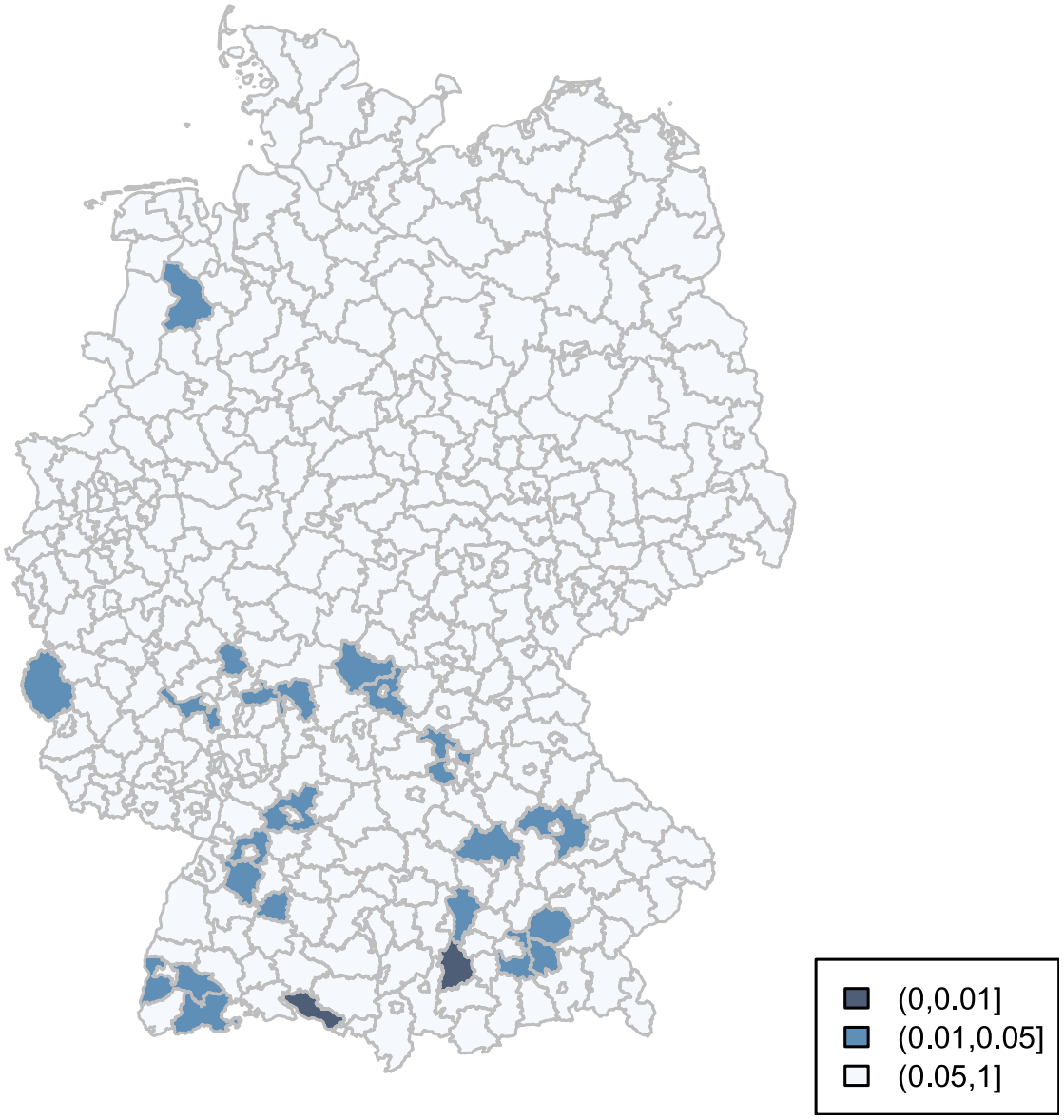

Figure 4.17: Choropleth map showing regions with log of SMR in the bottom percentile

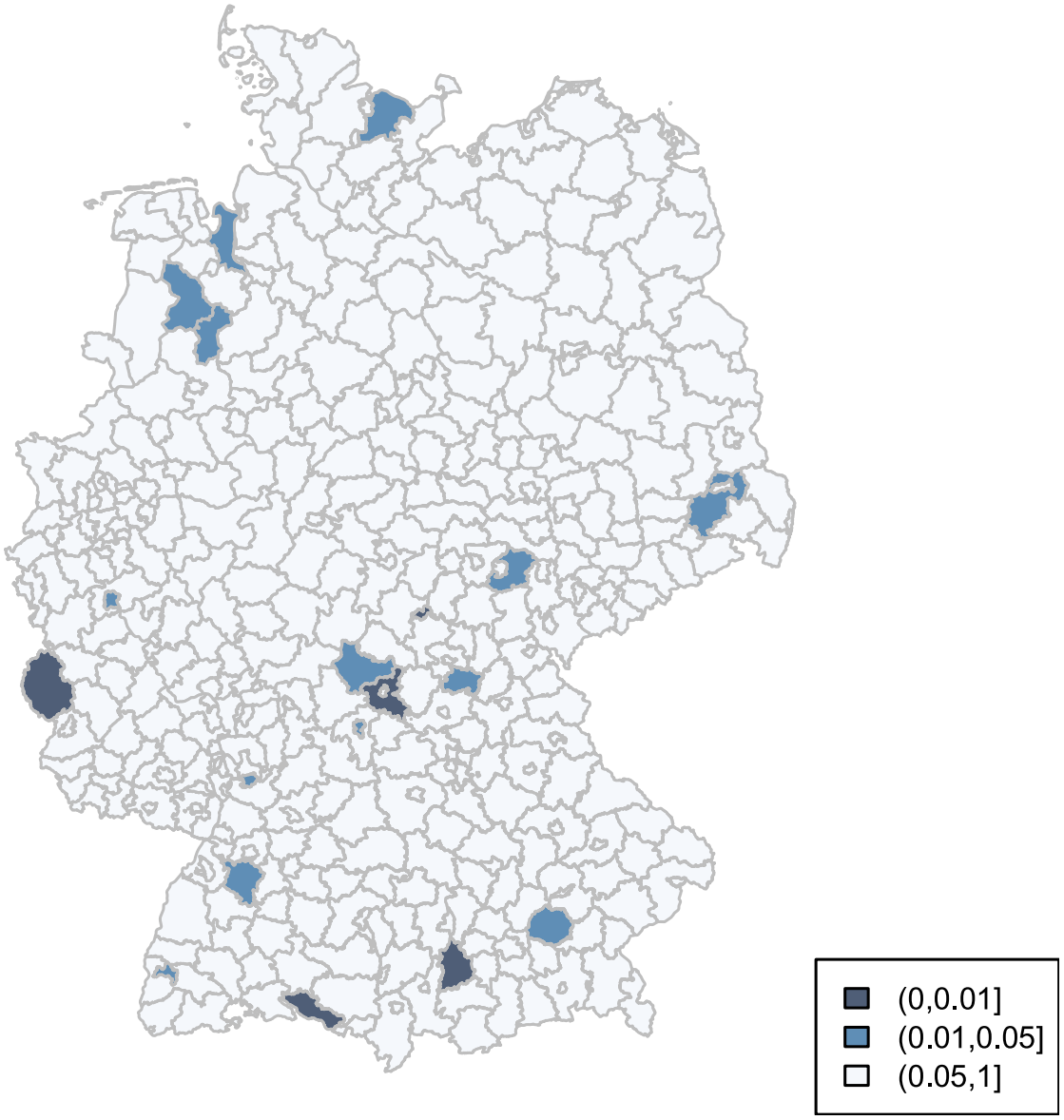

Figure 4.18: Choropleth map showing regions with residuals in the bottom percentile

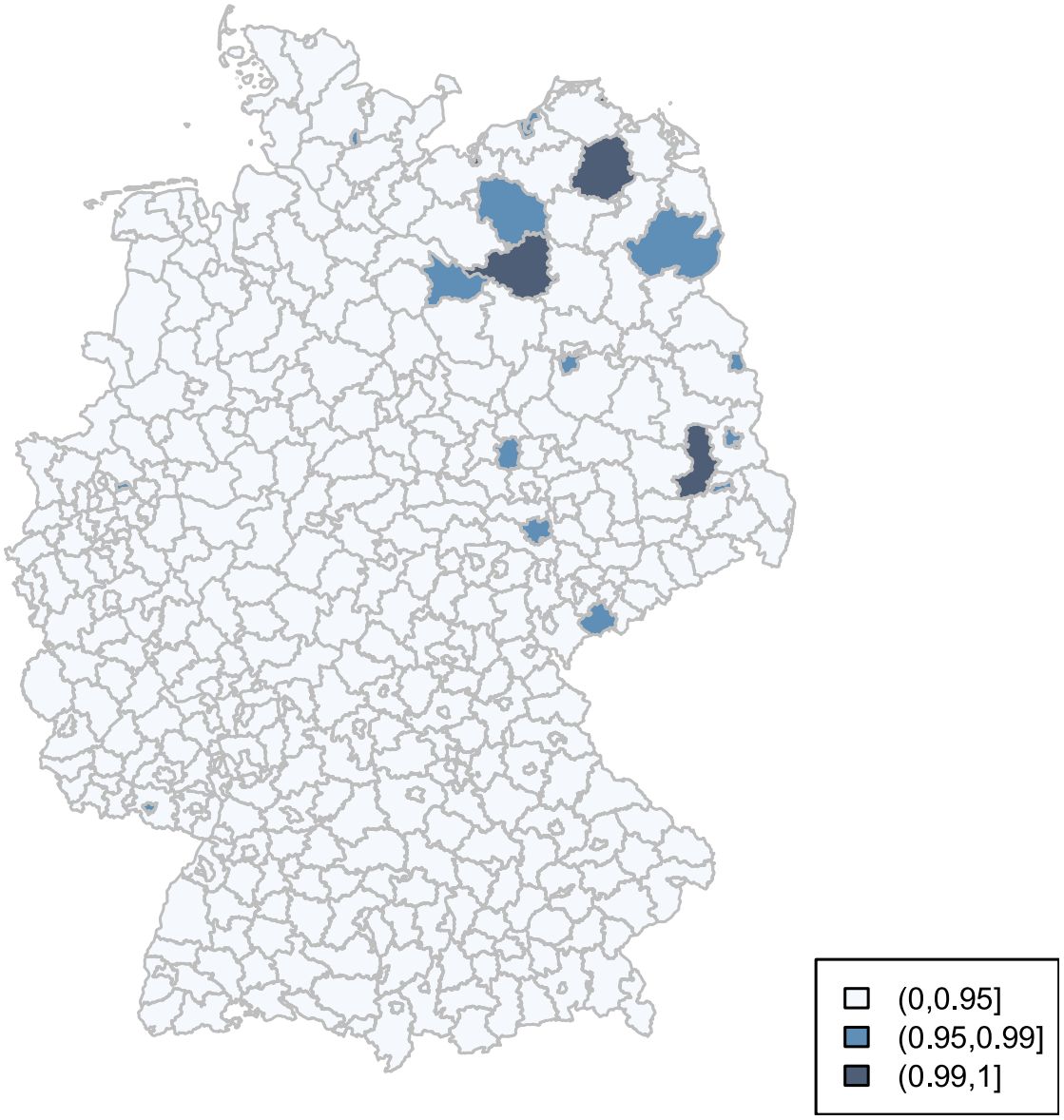

Figure 4.19: Choropleth map showing regions with log of SMR in the top percentile

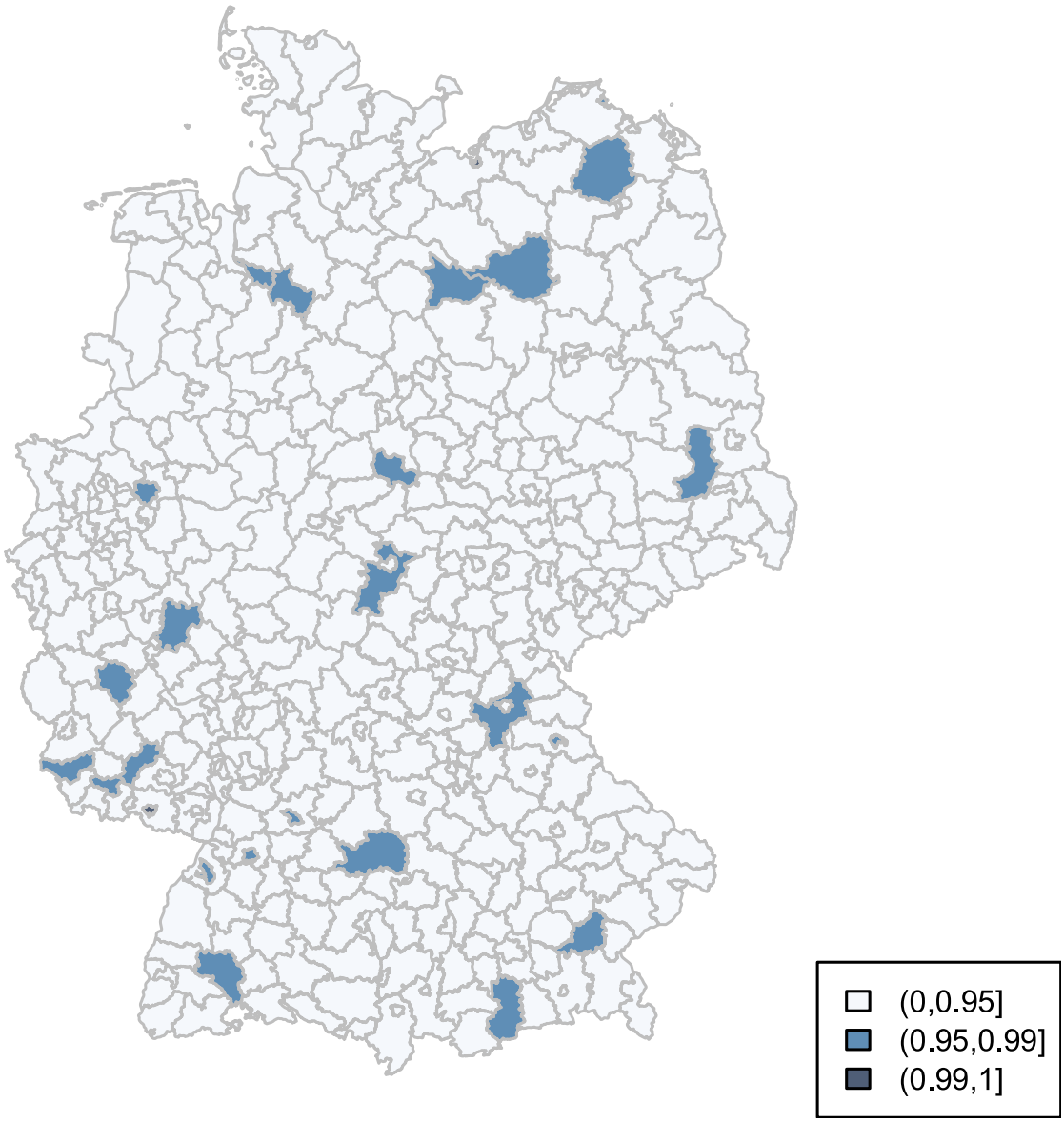

Figure 4.20: Choropleth map showing regions with residuals in the top percentile

Chapter 5

Conclusion

Our analysis revealed some interesting insights. Most importantly, we could show that mortality – as measured by the SMR – exhibits spatial dependency, warranting the need for spatial methods when analyzing mortality data aggregated at the regional level. We confirmed, the well established hypothesis that the income plays a role in overall mortality. We could not confirm the medical underspending hypothesis, however. On the contrary, regions with higher mortality also have more hospital beds, implying that municipalities adjust their spending to match the increased needs of their frailer population. The healthy migrant hypothesis – at least for the age group in question – could also be supported. Concerning the overall spatial trend we could show, that the North-to-South trend for all-cause mortality is not significant. Furthermore, a comparison of the coefficients, however, shows that in Germany rather a North-East-to-South-West trend exists than just a East-to-West trend in the

mortality differential.

When classifying regions into groups with unusual high or low mortality using the log of SMR, five regions are in top percentile of the distribution and two in the bottom percentile. Using the residuals of the regression model, however, to compare the fitted value for log of SMR with the observed values the picture is changing somewhat. Only two regions are now in the top percentile, but five in the bottom percentile. Hence, given the regression model – that controls for some explanatory variables and the spatial dependency of the dependent variable – less regions have an unusual high mortality compared with an univariate analysis, but more regions an unusual low mortality.

The findings have some limitations. First and foremost, it is a cross-sectional analysis and hence a snapshot in time. Although by using 5-year age groups – to ensure more stable mortality rates – a mortality analysis is more robust when the time dimension is modeled explicitly. Second, the analysis is at the aggregate level and ecological fallacy might reck its ugly head. Although, the causal relationships modeled have been extensively tested at the individual level, the findings should be considered with this limitation in mind.

From a demographic point of view, the estimated model could be extended in several ways. Instead of using both sexes for the calculation of the standard mortality ratio, it could have been split into males and females. But also additional variables could be included. Most notably, controlling for the settlement structure of the region could exercise an important influence; for example, ur-

ban versus rural regions. Another interesting variable could be the social composition of a region, such as the share of employees in the manufacturing industry, which is more hazardous for the age groups under consideration.

From a spatial statistic point of view, the analysis could be extend using a conditionally autoregressive (CAR) approach. The employed spatial error model assumes that the spatial random process simultaneously affects all regions. For our applications, however, it could be reasonable to assume that the spatial process differs locally. A CAR approach is doing this by conditioning only on neighboring regions.

Appendix A

Choropleth Maps of Independent Variables

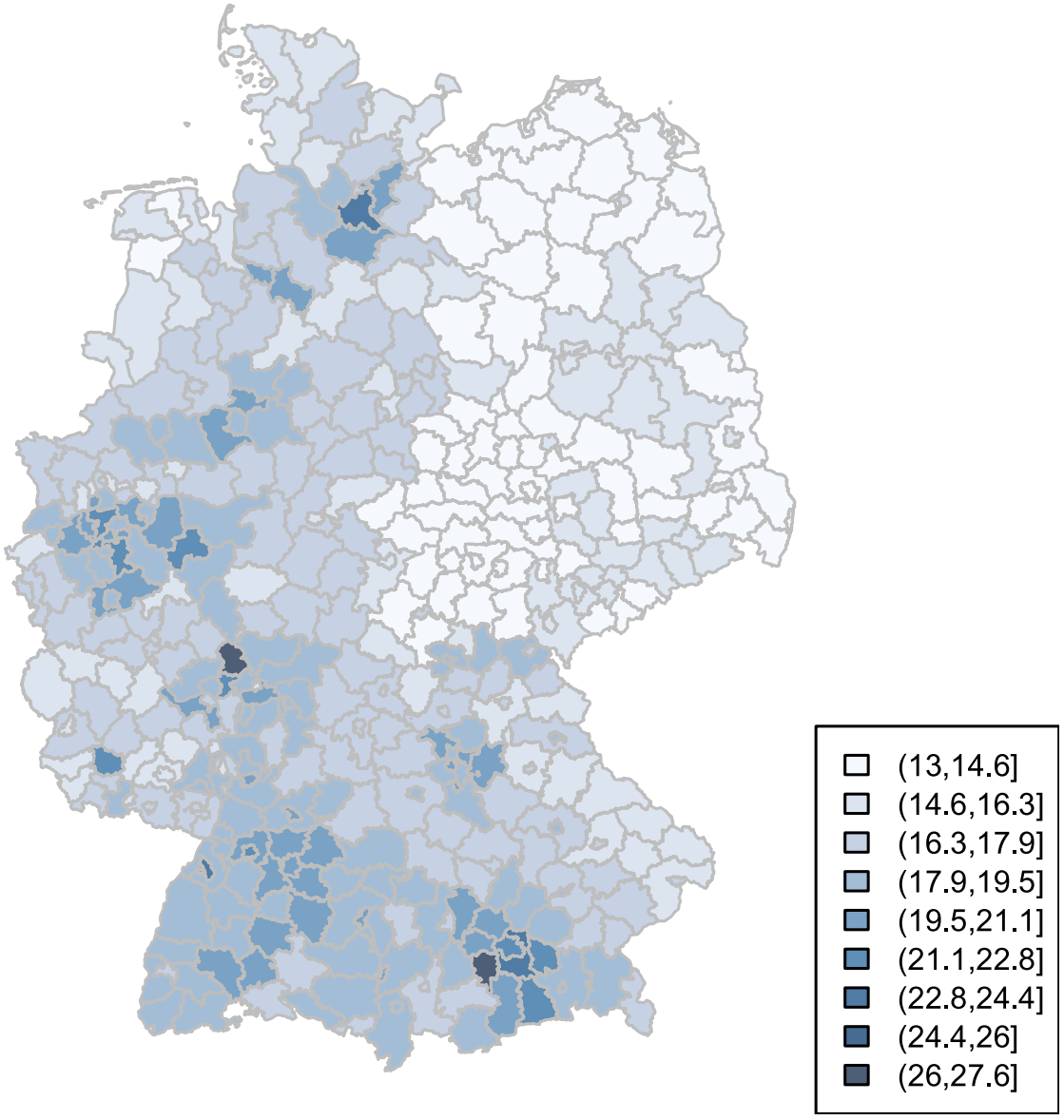

Figure A.1: Choropleth map of income per person in 1000 EUR

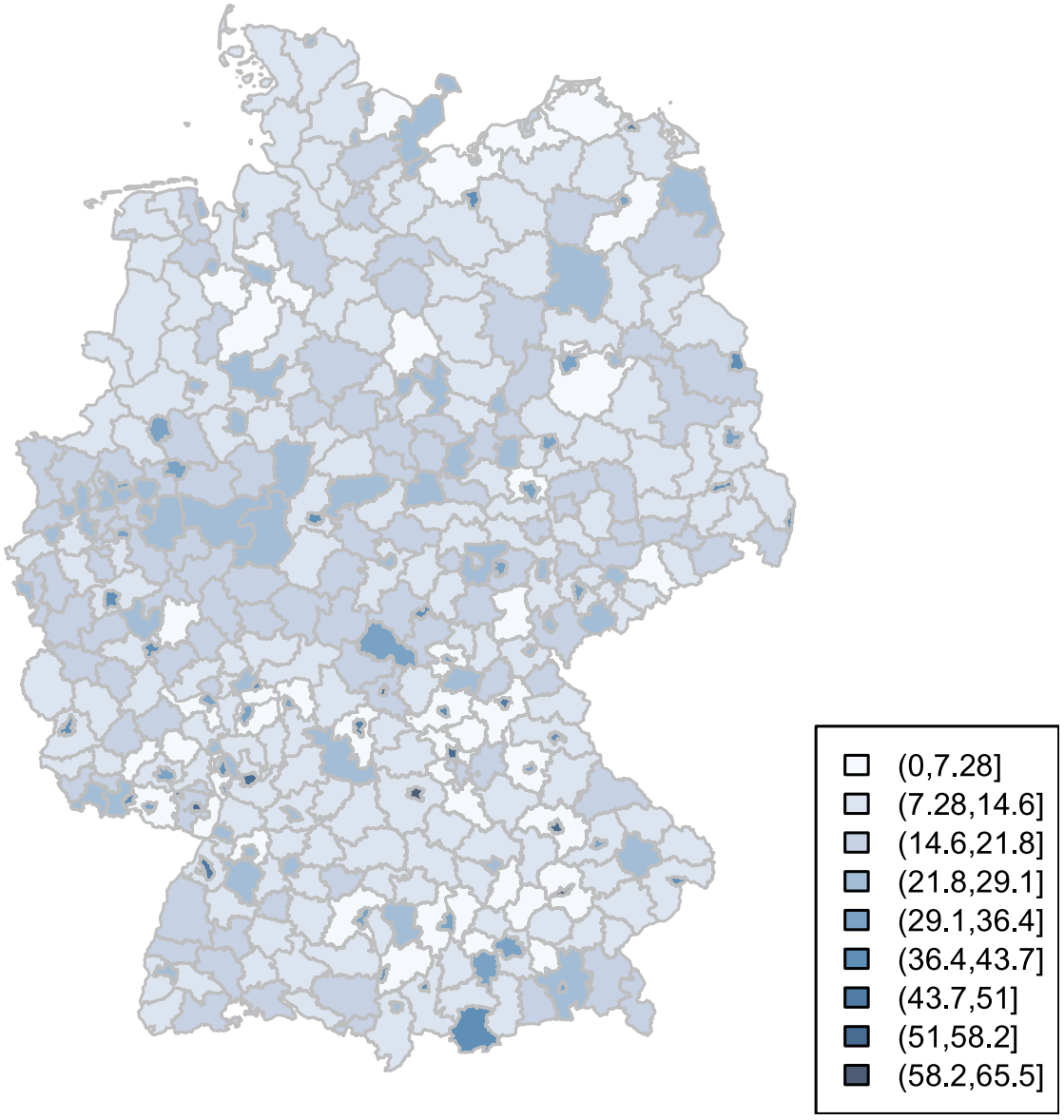

Figure A.2: Choropleth map of hospital beds per 1,000 residents

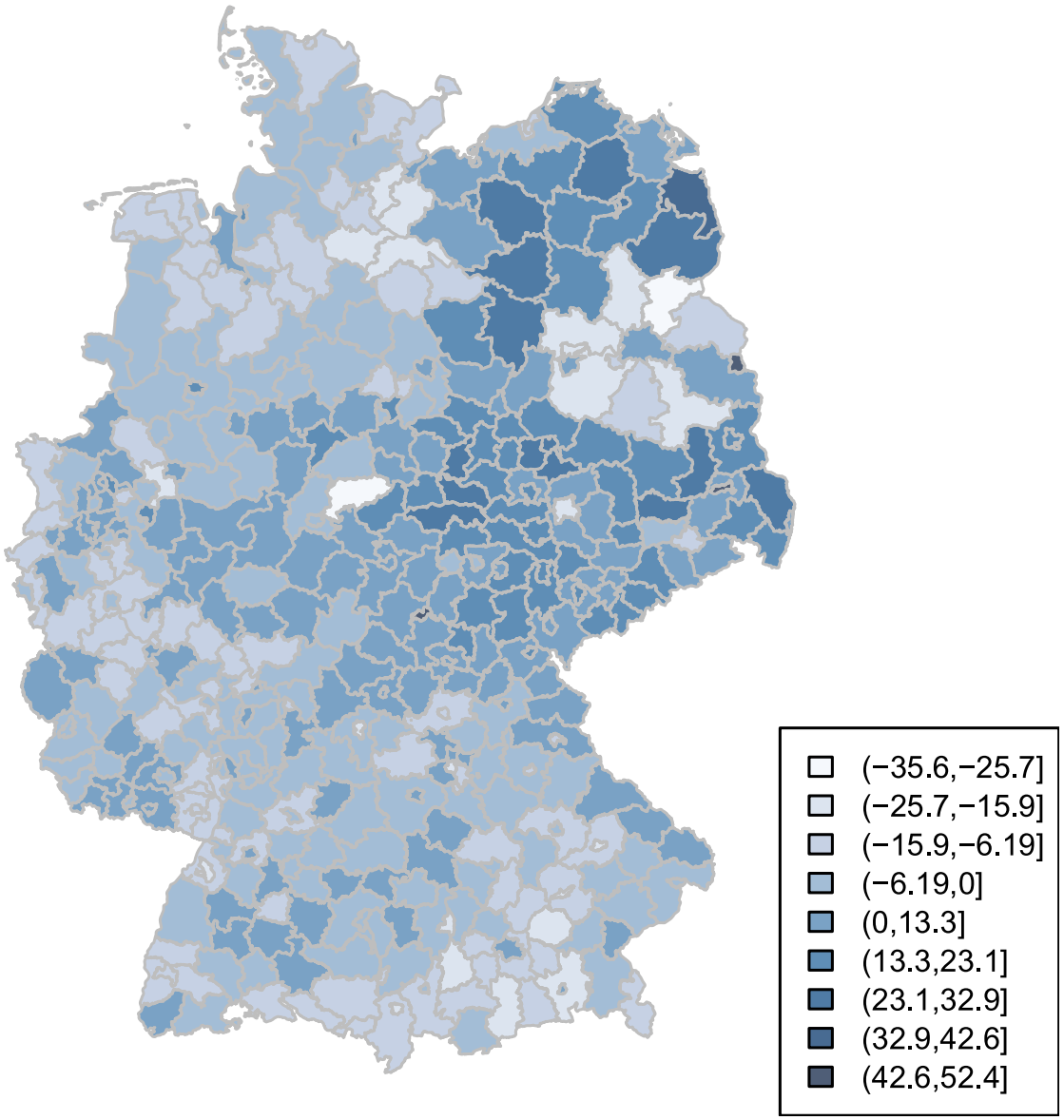

Figure A.3: Choropleth map of net migration (measured in out-migrants per 1,000 residents)

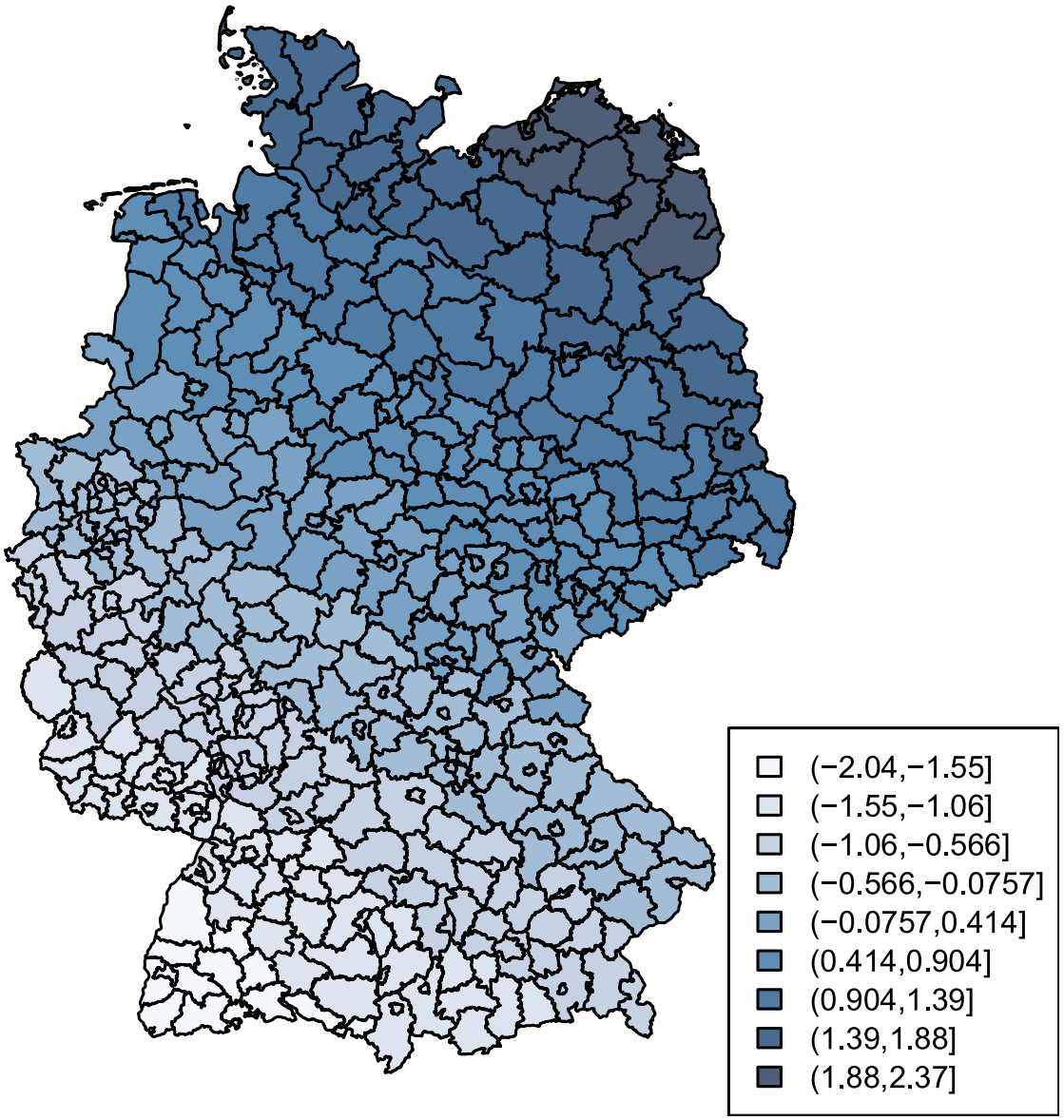

Figure A.4: Choropleth map of North/East-to-South /West-trend (standardized)

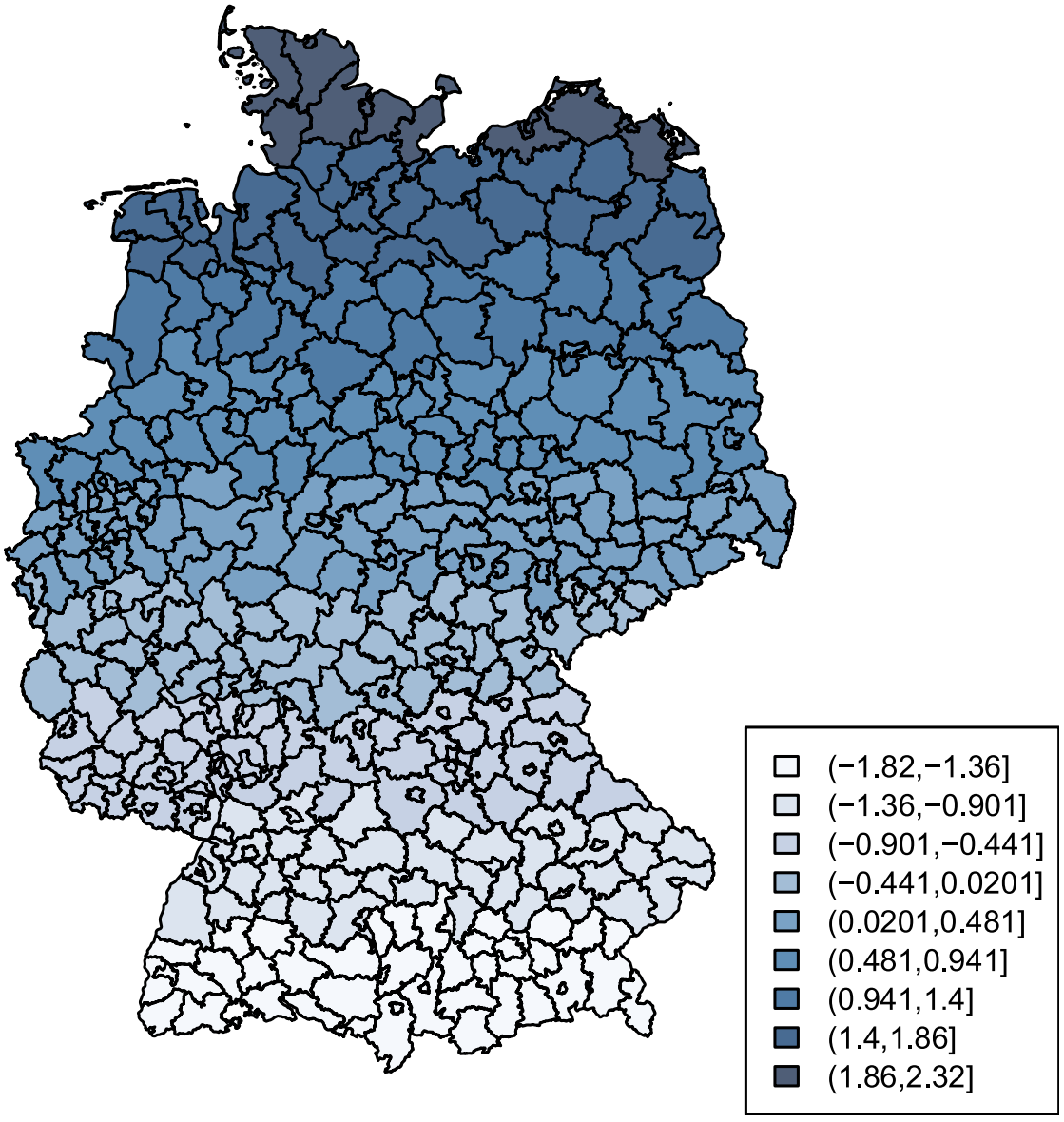

Figure A.5: Choropleth map of North-to-South-trend (standardized)

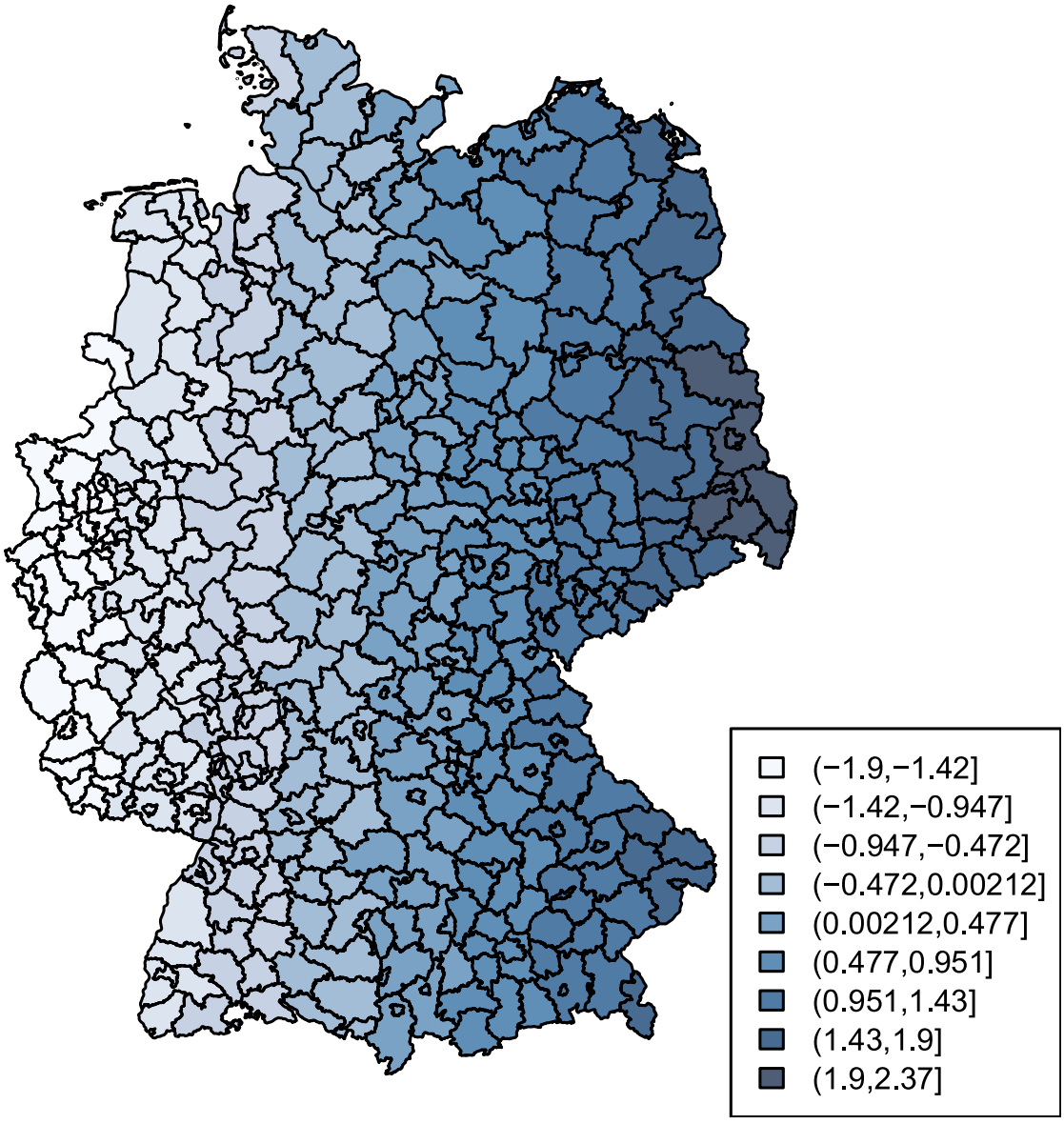

Figure A.6: Choropleth map of East-to-West-trend (standardized)

Appendix B

Overview of Regions

In this appendix we give an overview about the observations used. These are all the municipalities of Germany (except for the island of Rügen). We give the official name, administrative unit, the state to which they belong for the ease of identification. The number of observed deaths are for the age groups 20 to 49 from all causes for both sexes and are taken from the Federal Office of Statistics. The number of expected deaths have been calculated using an indirect standardization technique, namely applying the age adjusted national mortality rate to the population of the respective regions. The standard mortality ration can be computed by dividing the observed deaths by the expected deaths (for details see section 2.3).

List of all regions used in the analysis

ID	Name	Administrative Unit	Land	Obs. Deaths	Exp. Deaths
1	Nordfriesland	Kreis	SH	125	118.27
2	Ostholstein	Kreis	SH	169	146.78
4	Ostvorpommern	Kreis	MV	110	92.33
5	Pinneberg	Kreis	SH	228	222.13
6	Greifswald	kreisfreie Stadt	MV	53	40.37
7	Rostock	kreisfreie Stadt	MV	215	149.32
8	Steinburg	Kreis	SH	105	102.07
9	Dithmarschen	Kreis	SH	102	98.70
10	Rendsburg-Eckernförde	Kreis	SH	188	205.37
11	Lübeck	kreisfreie Stadt	SH	202	144.52
12	Parchim	Kreis	MV	138	88.99
13	Uecker-Randow	Kreis	MV	93	65.40
14	Bremerhaven	kreisfreie Stadt	HB	90	76.79
15	Harburg	Landkreis	NI	143	184.54
16	Rotenburg (Wümme)	Landkreis	NI	124	123.93
17	Uelzen	Landkreis	NI	80	69.10
18	Oldenburg	Landkreis	NI	86	98.52
19	Ostprignitz-Ruppin	Kreis	BR	113	91.97
20	Verden	Landkreis	NI	114	100.64
21	Osnabrück	Landkreis	NI	231	273.26
22	Berlin	Stadt u. Land	BE	2709	2426.39
23	Brandenburg a. d. H.	kreisfreie Stadt	BR	94	57.09
24	Osterode am Harz	Landkreis	NI	79	55.89
25	Göttingen	Landkreis	NI	173	204.78
26	Northeim	Landkreis	NI	108	104.03
27	Merseburg-Querfurt	Landkreis	ST	145	104.03
28	Aachen	kreisfreie Stadt	NW	163	180.93

Continued on Next Page...

ID	Name	Administrative Unit	Land	Obs. Death	Exp. Death
29	Aachen	Kreis	NW	210	219.95
30	Westerwaldkreis	Kreis	RP	159	150.46
31	Rhön-Grabfeld	Landkreis	BY	73	66.34
32	Schweinfurth	Landkreis	BY	55	90.40
33	Bayreuth	kreisfreie Stadt	BY	48	51.56
34	Tirschenreuth	Landkreis	BY	64	59.84
35	Haßberge	Landkreis	BY	58	70.39
36	Bamberg	Landkreis	BY	109	119.31
37	Rhein-Neckar-Kreis	Landkreis	BW	318	393.21
38	Bergstraße	Kreis	HE	165	194.69
39	Cham	Landkreis	BY	94	103.36
40	Karlsruhe	Landkreis	BW	267	319.08
41	Baden-Baden	Stadtkreis	BW	37	35.06
42	Ostalbkreis	Landkreis	BW	235	228.69
43	Rastatt	Landkreis	BW	134	170.36
44	Esslingen	Landkreis	BW	253	354.67
45	Alb-Donau-Kreis	Landkreis	BW	107	137.93
46	München	kreisfreie Stadt	BY	685	759.59
47	Rosenheim	Landkreis	BY	145	180.41
48	Konstanz	Landkreis	BW	172	197.48
50	Schleswig-Flensburg	Kreis	SH	163	146.48
51	Flensburg	kreisfreie Stadt	SH	68	58.52
52	Nordvorpommern	Kreis	MV	118	97.95
53	Plön	Kreis	SH	75	104.19
54	Kiel	kreisfreie Stadt	SH	193	174.74
55	Stralsund	kreisfreie Stadt	MV	79	44.73
56	Bad Doberan	Kreis	MV	110	106.40
57	Neumünster	kreisfreie Stadt	SH	81	54.40
58	Segeberg	Kreis	SH	185	196.02

Continued on Next Page...

ID	Name	Administrative Unit	Land	Obs. Death	Exp. Death
59	Demmin	Kreis	MV	136	74.84
60	Güstrow	Kreis	MV	125	89.74
61	Nordwestmecklenburg	Kreis	MV	108	105.42
62	Wismar	kreisfreie Stadt	MV	70	33.46
63	Stormarn	Kreis	SH	146	169.40
64	Cuxhaven	Landkreis	NI	151	144.70
65	Stade	Landkreis	NI	141	146.31
66	Herzogtum Lauenburg	Kreis	SH	135	139.48
67	Hamburg	Hansestadt	HH	1222	1206.43
68	Friesland	Landkreis	NI	73	73.66
69	Wittmund	Landkreis	NI	50	41.88
70	Müritz	Kreis	MV	65	57.24
71	Schwerin	kreisfreie Stadt	MV	106	73.92
72	Aurich	Landkreis	NI	152	138.94
73	Wilhelmshaven	kreisfreie Stadt	NI	65	58.21
74	Ludwigslust	Kreis	MV	148	111.05
75	Mecklenburg-Strelitz	Kreis	MV	90	72.63
76	Neubrandenburg	kreisfreie Stadt	MV	78	55.49
77	Wesermarsch	Landkreis	NI	50	67.11
78	Uckermark	Kreis	BR	174	115.19
79	Emden	kreisfreie Stadt	NI	33	35.50
80	Osterholz	Landkreis	NI	91	87.44
81	Lüneburg	Landkreis	NI	134	133.98
82	Leer	Landkreis	NI	131	119.04
83	Prignitz	Kreis	BR	121	72.54
84	Ammerland	Landkreis	NI	74	87.53
85	Oberhavel	Kreis	BR	183	169.48
86	Lüchow-Dannenberg	Landkreis	NI	51	35.32
87	Bremen	Freie Hansestadt	HB	475	368.10

Continued on Next Page...

ID	Name	Administrative Unit	Land	Obs. Death	Exp. Death
88	Soltau-Fallingbostel	Landkreis	NI	118	105.41
89	Oldenburg (Oldenburg)	kreisfreie Stadt	NI	102	122.38
90	Cloppenburg	Landkreis	NI	77	116.60
91	Emsland	Landkreis	NI	201	231.53
92	Delmenhorst	kreisfreie Stadt	NI	69	53.21
93	Barnim	Kreis	BR	167	149.49
94	Diepholz	Landkreis	NI	145	163.81
95	Stendal	Landkreis	ST	142	110.22
96	Altmarkkreis Salzwedel	Landkreis	ST	107	81.63
97	Celle	Landkreis	NI	146	129.95
98	Nienburg (Weser)	Landkreis	NI	102	93.32
99	Vechta	Landkreis	NI	73	99.71
100	Märkisch-Oderland	Kreis	BR	187	165.39
101	Gifhorn	Landkreis	NI	123	137.81
102	Havelland	Kreis	BR	132	131.09
103	Hannover, Region	Landkreis	NI	791	810.06
104	Grafschaft Bentheim	Landkreis	NI	89	91.06
105	Jerichower Land	Landkreis	ST	113	80.36
106	Potsdam-Mittelmark	Kreis	BR	141	174.58
107	Minden-Lübbecke	Kreis	NW	214	236.32
108	Ohrekreis	Landkreis	ST	110	100.24
109	Oder-Spree	Kreis	BR	170	156.47
110	Wolfsburg	kreisfreie Stadt	NI	81	79.58
111	Steinfurt	Kreis	NW	314	337.09
112	Helmstedt	Landkreis	NI	81	72.40
113	Schaumburg	Landkreis	NI	113	120.27
114	Potsdam	kreisfreie Stadt	BR	128	112.77
115	Peine	Landkreis	NI	88	100.18
116	Dahme-Spreewald	Kreis	BR	144	132.91

Continued on Next Page...

ID	Name	Administrative Unit	Land	Obs. Death	Exp. Death
117	Teltow-Fläming	Kreis	BR	163	136.00
118	Frankfurt (Oder)	kreisfreie Stadt	BR	78	51.99
119	Braunschweig	kreisfreie Stadt	NI	188	174.65
120	Osnabrück	kreisfreie Stadt	NI	111	121.93
121	Wolfenbüttel	Landkreis	NI	102	95.10
122	Hildesheim	Landkreis	NI	225	214.37
123	Hameln-Pyrmont	Landkreis	NI	135	110.39
124	Herford	Kreis	NW	177	184.64
125	Borken	Kreis	NW	226	275.39
126	Magdeburg	kreisfreie Stadt	ST	214	166.51
127	Salzgitter	kreisfreie Stadt	NI	73	72.85
128	Bördekreis	Landkreis	ST	78	63.85
129	Lippe	Kreis	NW	247	257.39
130	Anhalt-Zerbst	Landkreis	ST	79	60.06
131	Gütersloh	Kreis	NW	212	259.74
132	Bielefeld	kreisfreie Stadt	NW	223	222.41
133	Warendorf	Kreis	NW	179	207.33
134	Schönebeck	Landkreis	ST	77	57.41
135	Coesfeld	Kreis	NW	136	174.24
136	Münster	kreisfreie Stadt	NW	175	214.88
137	Halberstadt	Landkreis	ST	83	60.65
138	Spree-Neiße	Kreis	BR	159	115.54
139	Goslar	Landkreis	NI	125	107.15
140	Holzminden	Landkreis	NI	53	52.51
141	Wittenberg	Landkreis	ST	133	99.07
142	Aschersleben-Staßfurt	Landkreis	ST	109	76.11
143	Wernigerode	Landkreis	ST	81	72.00
144	Höxter	Kreis	NW	102	113.58
145	Quedlinburg	Landkreis	ST	67	57.35

Continued on Next Page...

ID	Name	Administrative Unit	Land	Obs. Death	Exp. Death
146	Kleve	Kreis	NW	225	225.34
147	Oberspreewald-Lausitz	Kreis	BR	175	103.68
148	Elbe-Elster	Kreis	BR	128	100.15
149	Bernburg	Landkreis	ST	75	51.89
150	Köthen	Landkreis	ST	72	53.00
151	Dessau	kreisfreie Stadt	ST	74	56.28
152	Paderborn	Kreis	NW	200	228.62
153	Cottbus	kreisfreie Stadt	BR	140	84.12
154	Recklinghausen	Kreis	NW	527	470.76
155	Wesel	Kreis	NW	332	357.59
156	Bitterfeld	Landkreis	ST	100	79.25
157	Hamm	kreisfreie Stadt	NW	136	124.74
158	Unna	Kreis	NW	313	311.83
159	Mansfelder Land	Landkreis	ST	89	78.64
160	Soest	Kreis	NW	255	227.15
161	Torgau-Oschatz	Kreis	SN	88	78.78
162	Saalkreis	Landkreis	ST	68	66.45
163	Delitzsch	Kreis	SN	113	101.15
164	Kassel	Kreis	HE	166	183.77
165	Nordhausen	Landkreis	TH	89	74.32
166	Bottrop	kreisfreie Stadt	NW	99	87.50
167	Sangerhausen	Landkreis	ST	53	50.40
168	Gelsenkirchen	kreisfreie Stadt	NW	254	183.00
169	Dortmund	kreisfreie Stadt	NW	546	381.95
170	Oberlausitzkreis	Kreis	SN	96	79.95
171	Eichsfeld	Landkreis	TH	86	88.8
172	Oberhausen	kreisfreie Stadt	NW	192	152.62
173	Herne	kreisfreie Stadt	NW	166	114.53
174	Duisburg	kreisfreie Stadt	NW	467	329.8

Continued on Next Page...

ID	Name	Administrative Unit	Land	Obs. Death	Exp. Death
175	Hochsauerlandkreis	Kreis	NW	188	198.98
176	Kamenz	Kreis	SN	100	120.59
177	Halle (Saale)	kreisfreie Stadt	ST	235	169.28
178	Essen	kreisfreie Stadt	NW	506	398.42
179	Bochum	kreisfreie Stadt	NW	280	279.09
180	Waldeck-Frankenberg	Kreis	HE	139	123.75
181	Ennepe-Ruhr-Kreis	Kreis	NW	270	249.87
182	Hoyerswerda	kreisfreie Stadt	SN	46	31.99
183	Märkischer Kreis	Kreis	NW	340	308.38
184	Mülheim a. d. Ruhr	kreisfreie Stadt	NW	116	118.18
185	Muldentalkreis	Kreis	SN	130	108.22
186	Riesa-Großenhain	Kreis	SN	113	89.86
187	Leipzig	kreisfreie Stadt	SN	435	356.03
188	Kyffhäuserkreis	Landkreis	TH	80	71.79
189	Viersen	Kreis	NW	205	230.42
190	Werra-Meißner-Kreis	Kreis	HE	88	81.46
191	Hagen	kreisfreie Stadt	NW	178	128.14
192	Krefeld	kreisfreie Stadt	NW	194	165.57
193	Mettmann	Kreis	NW	294	353.03
194	Bautzen	Kreis	SN	124	117.96
195	Kassel	kreisfreie Stadt	HE	186	136.05
196	Düsseldorf	kreisfreie Stadt	NW	432	387.66
197	Leipziger Land	Kreis	SN	155	118.27
198	Unstrut-Hainich-Kreis	Landkreis	TH	107	92.05
199	Neuss	Kreis	NW	272	322.10
200	Sömmerda	Landkreis	TH	78	64.93
201	Burgenlandkreis	Landkreis	ST	132	106.49
202	Wuppertal	kreisfreie Stadt	NW	285	236.86
203	Weißenfels	Landkreis	ST	87	59.41

Continued on Next Page...

ID	Name	Administrative Unit	Land	Obs. Death	Exp. Death
204	Schwalm-Eder-Kreis	Kreis	HE	132	143.89
205	Oberbergischer Kreis	Kreis	NW	192	208.86
206	Meißen	Kreis	SN	110	114.3
207	Olpe	Kreis	NW	83	104.44
208	Mönchengladbach	kreisfreie Stadt	NW	196	185.02
209	Döbeln	Kreis	SN	56	55.94
210	Görlitz	kreisfreie Stadt	SN	53	40.77
211	Remscheid	kreisfreie Stadt	NW	80	75.35
212	Solingen	kreisfreie Stadt	NW	101	112.40
213	Heinsberg	Kreis	NW	174	187.44
214	Löbau-Zittau	Kreis	SN	133	106.28
215	Dresden	kreisfreie Stadt	SN	342	346.44
216	RheinischBergischer Kreis	Kreis	NW	169	205.67
217	Wartburgkreis	Landkreis	TH	149	115.16
218	Siegen-Wittgenstein	Kreis	NW	198	209.26
219	Weimarer Land	Landkreis	TH	102	75.05
220	Gotha	Landkreis	TH	144	116.58
221	Mittweida	Kreis	SN	100	100.57
222	Altenburger Land	Landkreis	TH	108	83.08
223	Leverkusen	kreisfreie Stadt	NW	103	113.69
224	Sächsische Schweiz	Kreis	SN	124	104.12
225	Saale-Holzland-Kreis	Landkreis	TH	73	75.46
226	Hersfeld-Rotenburg	Kreis	HE	105	90.96
227	Köln	kreisfreie Stadt	NW	677	676.01
228	Erfurt	kreisfreie Stadt	TH	181	157.79
229	Weißeritzkreis	Kreis	SN	81	95.49
230	Erftkreis	Kreis	NW	319	335.04
231	Freiberg	Kreis	SN	94	109.44

Continued on Next Page...

ID	Name	Administrative Unit	Land	Obs. Death	Exp. Death
232	Eisenach	kreisfreie Stadt	TH	42	33.41
233	Düren	Kreis	NW	194	197.99
234	Weimar	kreisfreie Stadt	TH	59	48.63
235	Marburg-Biedenkopf	Kreis	HE	182	191.15
236	Jena	kreisfreie Stadt	TH	80	73.23
237	Gera	kreisfreie Stadt	TH	104	84.06
238	Greiz	Landkreis	TH	99	93.75
239	Rhein-Sieg-Kreis	Kreis	NW	378	441.72
240	Altenkirchen (Westerwald)	Kreis	RP	83	100.27
241	Chemnitzer Land	Kreis	SN	117	101.24
242	Ilm-Kreis	Landkreis	TH	107	91.78
243	Chemnitz	kreisfreie Stadt	SN	201	168.58
244	Lahn-Dill-Kreis	Kreis	HE	169	187.66
245	Zwickauer Land	Kreis	SN	107	98.37
246	Schmalkalden-Meiningen	Landkreis	TH	129	113.06
247	Vogelsbergkreis	Kreis	HE	96	88.15
248	Saalfeld-Rudolstadt	Landkreis	TH	117	99.02
249	Zwickau	kreisfreie Stadt	SN	90	71.98
250	Fulda	Kreis	HE	130	163.82
251	Mittlerer Erzgebirgskreis	Kreis	SN	70	67.81
252	Stollberg	Kreis	SN	62	67.84
253	Saale-Orla-Kreis	Landkreis	TH	72	76.16
254	Euskirchen	Kreis	NW	156	147.8
255	Bonn	kreisfreie Stadt	NW	177	219.79
256	Annaberg	Kreis	SN	87	62.71
257	Neuwied	Kreis	RP	134	135.52
258	Gießen	Kreis	HE	172	197.57
259	Vogtlandkreis	Kreis	SN	173	145.32

Continued on Next Page...

ID	Name	Administrative Unit	Land	Obs. Death	Exp. Death
260	Suhl	kreisfreie Stadt	TH	32	33.53
261	Aue-Schwarzenberg	Kreis	SN	148	98.90
262	Ahrweiler	Kreis	RP	74	94.74
263	Hildburghausen	Landkreis	TH	71	58.81
264	Limburg-Weilburg	Kreis	HE	117	130.39
265	Plauen	kreisfreie Stadt	SN	71	51.80
266	Sonneberg	Landkreis	TH	55	51.56
267	Kronach	Landkreis	BY	68	57.59
268	Wetteraukreis	Kreis	HE	175	224.39
269	Mayen-Koblenz	Kreis	RP	149	163.41
270	Main-Kinzig-Kreis	Kreis	HE	265	295.91
271	Bad Kissingen	Landkreis	BY	56	80.56
272	Hof	Landkreis	BY	91	78.80
273	Hochtaunuskreis	Kreis	HE	101	156.77
274	Rhein-Lahn-Kreis	Kreis	RP	86	95.29
275	Koblenz	kreisfreie Stadt	RP	95	73.27
276	Coburg	Landkreis	BY	63	70.57
277	Daun	Kreis	RP	41	45.91
278	Hof	kreisfreie Stadt	BY	43	32.03
279	Bitburg-Prüm	Kreis	RP	49	71.26
280	Coburg	kreisfreie Stadt	BY	35	29.61
281	Rheingau-Taunus-Kreis	Kreis	HE	109	137.44
282	Cochem-Zell	Kreis	RP	58	47.91
283	Kulmbach	Landkreis	BY	69	59.69
284	Rhein-Hunsrück-Kreis	Kreis	RP	84	80.00
285	Main-Spessart	Landkreis	BY	89	99.79
286	Wunsiedel i. Fichtelgebirge	Landkreis	BY	60	57.30
287	Frankfurt am Main	kreisfreie Stadt	HE	424	436.5

Continued on Next Page...

ID	Name	Administrative Unit	Land	Obs. Death	Exp. Death
288	Lichtenfels	Landkreis	BY	44	54.53
289	Main-Taunus-Kreis	Kreis	HE	133	158.89
290	Wiesbaden	kreisfreie Stadt	HE	167	172.51
291	Aschaffenburg	Landkreis	BY	89	133.50
292	Offenbach am Main	kreisfreie Stadt	HE	83	70.33
293	Offenbach	Kreis	HE	159	238.45
294	Bayreuth	Landkreis	BY	94	85.37
295	Bernkastel-Wittlich	Kreis	RP	80	83.64
296	Schweinfurt	kreisfreie Stadt	BY	30	32.28
297	Groß-Gerau	Kreis	HE	159	169.01
298	Mainz-Bingen	Kreis	RP	106	158.85
299	Mainz	kreisfreie Stadt	RP	122	124.64
300	Aschaffenburg	kreisfreie Stadt	BY	52	46.05
301	Darmstadt-Dieburg	Kreis	HE	182	214.70
302	Bad Kreuznach	Kreis	RP	114	113.61
303	Darmstadt	kreisfreie Stadt	HE	95	100.71
304	Würzburg	Landkreis	BY	97	127.82
305	Miltenberg	Landkreis	BY	77	94.91
306	Bamberg	kreisfreie Stadt	BY	40	50.31
307	Trier-Saarburg	Kreis	RP	87	107.65
308	Kitzingen	Landkreis	BY	56	68.52
309	Birkenfeld	Kreis	RP	70	65.10
310	Alzey-Worms	Kreis	RP	87	101.96
311	Forchheim	Landkreis	BY	78	90.16
312	Odenwaldkreis	Kreis	HE	71	69.68
313	Neustadt a.d. Waldnaab	Landkreis	BY	74	79.10
314	Trier	kreisfreie Stadt	RP	89	74.62
315	Würzburg	kreisfreie Stadt	BY	66	93.55
316	Main-Tauber-Kreis	Landkreis	BW	79	102.29

Continued on Next Page...

ID	Name	Administrative Unit	Land	Obs. Death	Exp. Death
317	Erlangen-Höchstadt	Landkreis	BY	66	103.05
318	Donnersbergkreis	Kreis	RP	73	60.94
319	Amberg-Sulzbach	Landkreis	BY	80	85.77
320	Neustadt a.d. Aisch-Bad Windsheim	Landkreis	BY	65	76.15
321	Worms	kreisfreie Stadt	RP	52	57.20
322	Nürnberger Land	Landkreis	BY	135	124.98
323	Weiden i.d. OPf.	kreisfreie Stadt	BY	41	30.28
324	Kusel	Kreis	RP	62	58.96
325	Neckar-Odenwald-Kreis	Landkreis	BW	85	114.85
326	Erlangen	kreisfreie Stadt	BY	69	71.39
327	St.Wendel	Landkreis	SL	73	73.13
328	Bad Dürkheim	Kreis	RP	88	101.38
329	Merzig-Wadern	Landkreis	SL	107	80.25
330	Ludwigshafen	Kreis	RP	91	113.03
331	Kaiserslautern	Kreis	RP	80	83.67
332	Schwandorf	Landkreis	BY	116	112.92
333	Mannheim	Stadtkreis	BW	229	196.99
334	Frankenthal (Pfalz)	kreisfreie Stadt	RP	36	31.31
335	Fürth	Landkreis	BY	58	91.33
336	Ludwigshafen am Rhein	kreisfreie Stadt	RP	113	100.77
337	Fürth	kreisfreie Stadt	BY	79	79.34
338	Nürnberg	kreisfreie Stadt	BY	368	320.38
339	Saarlouis	Landkreis	SL	172	155.05
340	Amberg	kreisfreie Stadt	BY	34	32.15
341	Kaiserslautern	kreisfreie Stadt	RP	94	69.95
342	Ansbach	Landkreis	BY	115	140.00
343	Heidelberg	Stadtkreis	BW	79	103.62
344	Neunkirchen	Landkreis	SL	144	106.90

Continued on Next Page...

ID	Name	Administrative Unit	Land	Obs. Death	Exp. Death
345	Neumarkt i.d. OPf.	Landkreis	BY	102	102.05
346	Hohenlohekreis	Landkreis	BW	63	82.50
347	Saarpfalz-Kreis	Landkreis	SL	107	116.07
348	Neustadt a. d. Weinstraße	kreisfreie Stadt	RP	36	39.51
349	Schwäbisch Hall	Landkreis	BW	122	140.59
350	Roth	Landkreis	BY	79	98.69
351	Heilbronn	Landkreis	BW	163	240.40
352	Stadtverband Saarbrücken	Stadtverband	SL	330	244.65
353	Südwestpfalz	Kreis	RP	66	80.64
354	Schwabach	kreisfreie Stadt	BY	29	27.61
355	Speyer	kreisfreie Stadt	RP	43	35.77
356	Südliche Weinstraße	Kreis	RP	78	85.26
357	Ansbach	kreisfreie Stadt	BY	31	27.08
358	Zweibrücken	kreisfreie Stadt	RP	28	25.65
359	Germersheim	Kreis	RP	99	96.09
360	Landau in der Pfalz	kreisfreie Stadt	RP	36	31.97
361	Pirmasens	kreisfreie Stadt	RP	44	28.55
362	Regensburg	Landkreis	BY	104	149.04
363	Heilbronn	Stadtkreis	BW	87	74.56
364	Weißenburg-Gunzenhausen	Landkreis	BY	81	69.68
365	Regen	Landkreis	BY	72	62.43
366	Straubing-Bogen	Landkreis	BY	63	78.81
367	Karlsruhe	Stadtkreis	BW	174	197.64
368	Eichstätt	Landkreis	BY	67	96.62
369	Rems-Murr-Kreis	Landkreis	BW	228	288.75
370	Regensburg	kreisfreie Stadt	BY	89	96.02

Continued on Next Page...

ID	Name	Administrative Unit	Land	Obs. Death	Exp. Death
371	Ludwigsburg	Landkreis	BW	289	350.98
372	Enzkreis	Landkreis	BW	89	143.07
373	Donau-Ries	Landkreis	BY	76	98.38
374	Kehlheim	Landkreis	BY	78	86.87
375	Freyung-Grafenau	Landkreis	BY	59	65.65
376	Deggendorf	Landkreis	BY	85	92.36
377	Pforzheim	Stadtkreis	BW	89	74.47
378	Straubing	kreisfreie Stadt	BY	37	31.12
379	Böblingen	Landkreis	BW	181	255.98
380	Stuttgart	Stadtkreis	BW	327	371.27
381	Calw	Landkreis	BW	80	114.89
382	Neuburg-Schrobenhausen	Landkreis	BY	53	69.47
383	Ingolstadt	kreisfreie Stadt	BY	70	80.06
384	Pfaffenhofen a.d. Ilm	Landkreis	BY	74	91.54
385	Dingolfing-Landau	Landkreis	BY	75	70.98
386	Heidenheim	Landkreis	BW	102	92.89
387	Göppingen	Landkreis	BW	141	174.26
388	Landshut	Landkreis	BY	93	118.01
389	Passau	Landkreis	BY	118	145.79
390	Dillingen a.d. Donau	Landkreis	BY	60	71.48
391	Ortenaukreis	Landkreis	BW	265	307.68
392	Freudenstadt	Landkreis	BW	90	89.40
393	Augsburg	Landkreis	BY	135	183.3
394	Tübingen	Landkreis	BW	115	167.59
395	Aichach-Friedberg	Landkreis	BY	64	99.16
396	Rottal-Inn	Landkreis	BY	77	89.98
397	Freising	Landkreis	BY	90	122.79
398	Passau	kreisfreie Stadt	BY	44	35.99

Continued on Next Page...

ID	Name	Administrative Unit	Land	Obs. Death	Exp. Death
399	Reutlingen	Landkreis	BW	179	199.04
400	Landshut	kreisfreie Stadt	BY	44	41.20
401	Günzburg	Landkreis	BY	85	89.85
402	Neu-Ulm	Landkreis	BY	101	114.92
403	Ulm	Stadtkreis	BW	65	79.17
404	Augsburg	kreisfreie Stadt	BY	152	166.18
405	Dachau	Landkreis	BY	80	100.06
406	Erding	Landkreis	BY	61	96.73
407	Mühldorf a. Inn	Landkreis	BY	91	82.97
408	Zollernalbkreis	Landkreis	BW	119	137.79
409	Rottweil	Landkreis	BW	77	101.59
410	Altötting	Landkreis	BY	78	80.00
411	München	Landkreis	BY	140	212.58
412	Fürstenfeldbruck	Landkreis	BY	109	143.03
413	Biberach	Landkreis	BW	104	142.10
414	Sigmaringen	Landkreis	BW	85	99.69
415	Emmendingen	Landkreis	BW	102	119.39
416	Landsberg am Lech	Landkreis	BY	45	86.50
417	Ebersberg	Landkreis	BY	64	92.58
418	Schwarzwald-Baar-Kreis	Landkreis	BW	162	144.47
419	Tuttlingen	Landkreis	BW	68	94.14
420	Traunstein	Landkreis	BY	98	122.89
421	Starnberg	Landkreis	BY	71	89.07
422	Breisgau-Hochschwarzwald	Landkreis	BW	128	184.07
423	Ostallgäu	Landkreis	BY	83	98.86
424	Freiburg im Breisgau	Stadtkreis	BW	115	162.43
425	Memmingen	kreisfreie Stadt	BY	30	26.61
426	Unterallgäu	Landkreis	BY	77	99.27

Continued on Next Page...

ID	Name	Administrative Unit	Land	Obs. Death	Exp. Death
427	Ravensburg	Landkreis	BW	162	201.21
428	Bad Tölz- Wolfratshausen	Landkreis	BY	83	86.80
429	Berchtesgadener Land	Landkreis	BY	62	65.49
430	Weilheim-Schongau	Landkreis	BY	77	95.83
431	Miesbach	Landkreis	BY	48	67.97
432	Kaufbeuren	kreisfreie Stadt	BY	27	28.89
433	Rosenheim	kreisfreie Stadt	BY	40	40.31
434	Bodenseekreis	Landkreis	BW	84	143.67
435	Lörrach	Landkreis	BW	133	160.36
436	Oberallgäu	Landkreis	BY	93	107.76
437	Waldshut	Landkreis	BW	82	117.03
439	Kempten (Allgäu)	kreisfreie Stadt	BY	43	39.14
440	Garmisch-Partenkirchen	Landkreis	BY	66	59.13
441	Lindau (Bodensee)	Landkreis	BY	56	52.96

Appendix C

Used Software

The statistical computations and map drawings for this thesis were exclusively made using R, a freely available software package. R lives from the contribution of many dedicated statisticians that allows R to be a flexible, up-to-date program for almost all applications. In this appendix an overview is given about the most important libraries – and their respective authors – that were used in creating the maps and conducting the statistical analysis.

- *spdep – Spatial dependence: weighting schemes, statistics and models* (R package version 0.4-13)
  by Roger Bivand and with contributions by Luc Anselin and Olaf Berke and Andrew Bernat and Marilia Carvalho and Yongwan Chun and Carsten Dormann and Stéphane Dray and Rein Halbersma and Nicholas Lewin-Koh and Jielai Ma and Giovanni Millo and Werner Mueller and Hisaji Ono and Pedro Peres-Neto and Markus Reder and Michael Tiefelsdorf and Danlin Yu.

- *maptools – Tools for reading and handling spatial objects* (R package version 0.7-4)

by Nicholas J. Lewin-Koh and Roger Bivand and contributions by Edzer J. Pebesma and Eric Archer and Stéphane Dray and David Forrest and Patrick Giraudoux and Duncan Golicher and Virgilio Gómez Rubio and Patrick Hausmann and Thomas Jagger and Sebastian P. Luque and Don MacQueen and Andrew Niccolai and Tom Short

- *DCluster – Detecting clusters of disease with R* (R package version 0.2)
  by V. Gomez-Rubio; J. Ferrandiz-Ferragud; A. Lopez-Qualez

- *RColorBrewer – ColorBrewer palettes* (R package version 1.0-2)
  by Erich Neuwirth

- *RODBC – ODBC Database Access* (R package version 1.2-3)
  by Originally Michael Lapsley and from Oct 2002 B. D. Ripley

Bibliography

Luc Anselin. *A workbook for using SpaceStat in the ananlysis of spatial data.* Working paper, Urbana–Champaign, 1992.

Luc Anselin. Spatial Econometrics. In Badi H. Baltagi, editor, *A companion to theoretical econometrics*, pages 310–330. Blackwell, Malden, 2003.

Luc Anselin. *Spatial regression.* Working paper, Urbana–Champaign, 2006.

Sudipto Banerjee, Bradley P. Carlin, and Alan E. Gelfand. *Hierarchical modeling and analysis for spatial data.* Chapman & Hall, Boca Raton, 2004.

Bundesamt für Kartographie. Vektordaten für die BRD: Verwaltungsgrenzen, 2004.

Bundesamt für Statistik. DVD Statistik regional, 2007.

G. Caselli, E. Barbi, C. Tomassini, S. Francisci, R. M. Lipsi, and D. Pierannunzio. Analysis of Sardinian Mortality. Working paper MPIDR. Rostock, 2002.

Guangqing Chi and Jun Zhu. Spatial regression models for demographic analysis. *Population Research and Policy Review*, 27(1):17–42, 2008.

Chin Long Chiang. *The life table and its applications.* Krieger, Malabar, FL, 1984.

Pavel Čižek, Wolfgang Härdle, and Jürgen Symanzik. Spatial Statistics. In Wolfgang Härdle, Yuichi Mori, and Philippe Vieu, editors, *Statistical methods for biostatistics and related fields*, pages 285–304. Springer, Berlin, 2007.

Ansley Johnson Coale, Paul Demeny, and Barbara Vaughan. *Regional model life tables and stable populations.* Academic Press, New York, 2nd edition, 1983.

Noel A. C. Cressie. *Statistics for spatial data.* Wiley, New York, 1993.

Jürgen Cromm and Rembrandt D. Scholz, editors. *Regionale Sterblichkeit in Deutschland.* WiSoMed, Göttingen, 2002.

David Cutler, Angus Deaton, and Adriana Lleras Muney. The determinants of mortality. *Journal of Economic Perspectives*, 20(3):97–120, 2006.

David Darmofal. *Spatial Econometrics and Political Science*. Society for Political Methodology, 2007.

V. Gómez Rubio, J. Ferrándiz, and A. López. Detecting disease clusters with R. In K. Hornik, F. Leisch, and A. Zeilis, editors, *Proceedings of the 3rd international workshop on distributed statistical computing*, pages 15–29. TU Wien, 2003.

Sander Greenland. Divergent biases in ecologic and individual–level studies. *Statistics in Medicine*, 11(9):1209–1223, 1992.

Robert P. Haining. *Spatial data analysis: Theory and practice*. Cambridge Univ. Press, Cambridge, 2005.

Francesco Lagona and Elisabeth Barbi. *Spatial demography*. Lecture notes, Rome, 2006.

Andrew B. Lawson. *Statistical methods in spatial epidemiology*. Wiley, Chichester, 2001.

James P. LeSage. *The theory and practice of spatial econometrics*. Manuscript, Toledo, 1999.

Roger J. Marshall. Mapping disease and mortality rates using empirical bayes estimators. *Applied Statistics*, 40(2):283–294, 1991.

Keith Ord. Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, 70(349):120–126, 1975.

Stuart J. Pocock, Derek G. Cook, and Shirley A.A. Beresford. Regression of area mortality rates on explanatory variables: What weighting is appropriate? *Applied Statistics*, 30(3):286–295, 1981.

Samuel H. Preston, Patrick Heuveline, and Michel Guillot. *Demography: Measuring and modeling population processes*. Blackwell, Oxford, 2005.

Roland Rau. *Seasonality in human mortality*. Springer, Berlin, 2007.

Henry S. Shryock and Jacob S. Siegel. *The methods and materials of demography*. Academic Press, San Diego, 1988.

A. Roger Thatcher, Väinö Kannisto, and James W. Vaupel. *The force of mortality at ages 80 to 120*. Odense University Press, Odense, Denmark, 1998.

Edward Rolf Tufte. *The visual display of quantitative information.* Graphics Press, Cheshire, 1999.

Jacques Vallin and Graziella Caselli. Cohort Life Table. In Graziella Caselli, Jacques Vallin, Guillaume J. Wunsch, and Daniel Courgeau, editors, *Demography: Analysis and synthesis*, volume 1, pages 103–128. Elsevier, 2006.

Melanie M. Wall. A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference*, 121(2):311–324, 2004.

Jeffrey M. Wooldridge. *Introductory econometrics: A modern approach.* Thomson, Mason, 2 edition, 2005.

Declaration of Authorship

I hereby solemnly declare that I have authored this master thesis. All sources and material used for this thesis are duely referenced and all quotes are clearly marked as such. Berlin, March 10th, 2008

Stefan K. Lhachimi